Benzaid, Chafika; Taleb, Tarik

AI-driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions

Published in:
IEEE NETWORK

DOI:
10.1109/MNET.001.1900252

Published: 01/02/2020

Document Version
Peer reviewed version

Please cite the original version:
Benzaid, C., & Taleb, T. (2020). AI-driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions. *IEEE NETWORK*, *34*(2), 186 - 194.
https://doi.org/10.1109/MNET.001.1900252

AI-driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions

Chafika Benzaïd* and Tarik Taleb†
*†Aalto University, Espoo, Finland
†University of Oulu, Oulu, Finland
†Sejong University, Seoul, South Korea
Email: *chafika.benzaid@aalto.fi, †tarik.taleb@aalto.fi

***Abstract*—The foreseen complexity in operating and managing 5G and beyond networks has propelled the trend toward closed-loop automation of network and service management operations. To this end, the ETSI Zero-touch network and Service Management (ZSM) framework is envisaged as a next-generation management system that aims to have all operational processes and tasks executed automatically, ideally with 100% automation. Artificial Intelligence (AI) is envisioned as a key enabler of self-managing capabilities, resulting in lower operational costs, accelerated time-to-value and reduced risk of human error. Nevertheless, the growing enthusiasm for leveraging AI in a ZSM system should not overlook the potential limitations and risks of using AI techniques. The current paper aims to introduce the ZSM concept and point out the AI-based limitations and risks that need to be addressed in order to make ZSM a reality.**

***Index Terms*—5G, ZSM, Artificial Intelligence, Machine Learning, and Network Management.**

I. Introduction

THE fifth generation of mobile communication networks (5G) comes as an answer to the foreseen demands of high traffic volume, massive number of connected devices with diverse service requirements, better quality of user experience (QoE) and better affordability by further reducing costs. Compared to 4G LTE, it is envisioned that the upcoming 5G networks will bring near-to-zero round-trip latency, 10 times higher data rate, "five $9s$" availability, almost 100% coverage, up to 90% reduction in energy usage and $10-100$ more connected devices [1]. Thanks to its key network capabilities, 5G networks will be a pivotal enabler of emerging usage scenarios and applications. Indeed, three usage scenarios are envisaged by ITU IMT-2020, namely enhanced mobile broadband (eMBB), addressing the human-centric use cases for access to multimedia content, services and data; ultra-reliable and low-latency communications (URLLC) with stringent requirements in terms of latency and reliability; and massive machine type communications (mMTC) for a very large number of connected devices typically transmitting a relatively low volume of non-delay sensitive data.

To leverage the promising 5G capabilities in order to fulfill the very disparate and challenging requirements of those future use cases, 5G networks are being conceived as extremely flexible, highly programmable and holistically-managed infrastructures that are service- and context-aware [1]. To this end, emerging technologies and concepts such as Software Defined Networking (SDN), Network Function Virtualization (NFV), Multi-access Edge Computing (MEC) and Network Slicing are identified to play a key role in the design of 5G network architecture. The use of these technologies will unlock new business models, including multi-domain, multi-service, multi-tenancy models, to support new markets. Meanwhile, the increase in performance, flexibility and cost efficiency, coupled with the imposed agility and cooperation across domains, are expected to result in unprecedented complexity in operating and managing 5G networks. Thus, traditional service and network management solutions may not be sufficient, making *closed-loop automation* of management operations an inevitability. The shift to management automation will boost the flexibility and efficiency of service delivery and reduce the OPerating EXpenses (OPEX) through self-managing capabilities (e.g., self-configuration, self-healing, self-optimization, and self-protecting). Being aware of the importance of management automation in 5G, the topic has gained much attention from both research community and Standards Developing Organizations (SDOs). However, most efforts have revolved around enabling automation in a single domain. To meet the challenging performance requirements of the various 5G usage scenarios, an End-to-End (E2E) service and network management automation across multiple domains is needed. To this aim, ETSI established the Zero Touch network and Service Management Industry Specification Group (ZSM ISG) in 2017. A primary goal of the ETSI ZSM ISG is to specify an end-to-end network and service management reference architecture enabling agile, efficient and qualitative management and automation of emerging and future networks and services. The ZSM framework is envisaged as a next-generation management system that aims to have all operational processes and tasks (e.g., planning and design, delivery, deployment, provisioning, monitoring and optimization) executed automatically, ideally with 100% automation and without human intervention. Artificial Intelligence (AI), supported by Machine Learning (ML) and Big Data analytics techniques, is envisioned as a key enabler of fully autonomous networks. Tractica foresees that the spending on AI-driven network management software will increase from $\$23$ million in 2018 to more than $\$1.9$ billion in 2021, with annual spend

to reach $7.4 billion by 2025. AI plays an important role in empowering self-managing functionalities, resulting in lower operational costs, accelerated time-to-value and reduced risk of human error. Nevertheless, the growing enthusiasm for leveraging AI in a ZSM system should not overlook the potential limitations and risks of using AI techniques.

The current article aims to introduce the ZSM concept and point out the AI-based limitations that need to be addressed in order to make ZSM a reality. We first present research contributions and SDOs initiatives that could be leveraged by ZSM. Then, we describe the ZSM reference architecture. Following that, we discuss the limitations and security risks related to the use of AI techniques in ZSM. This leads us to highlight some future research directions to tackle the identified issues. Finally, we conclude the article.

II. Research and Standardization Work Relevant to ZSM

In this section, we discuss the existing research contributions and standardization initiatives that adopt AI techniques to enable intelligent and automated service and network management in next-generation networks.

A. Academic Research Work

Over the past few decades, AI/ML techniques have been leveraged to intelligently perform a variety of networking operations in future networks, ranging from management to maintenance and protection. Fadlullah *et al.* [2] surveyed the works on Deep Learning (DL) applications for various traffic control aspects, such as network traffic classification, network flow prediction, mobility prediction, Cognitive Radio Networks (CRNs), and Self-Organized Networks (SONs). The authors demonstrated the effectiveness of a deep-learning routing approach compared to a conventional routing strategy in a wireless mesh backbone network.

Authors in [3] proposed a closed-loop solution for network slicing where traffic forecasting information is ingested by an admission control engine to maximize the number of granted network slice requests while meeting the slice Service Level Agreements (SLAs) guarantees. A slice scheduling module is in charge of provisioning physical resources to admitted slice requests and reporting SLA violations to the forecasting module in order to correct the foreseen traffic load. The per-slice traffic prediction is based on past traffic and user mobility using the Holt-Winters time-series forecasting model. The admission control problem is formulated as a geometric knapsack problem and is solved with a simulated annealing-based heuristic.

Martin *et al.* [4] developed a network resource allocator system that fosters self-configuration, self-optimization and self-healing capabilities by means of ML, SDN and NFV technologies. The system enables QoE-aware autonomous network management which dynamically and proactively provisions a network topology to adapt to changing demands of media services. To this end, a ML engine is integrated into SDN controller to forecast traffic load and corresponding KPIs, and foresee the network topology to be setup in line with SLA requirements. The ML engine comprises three modules, namely: (i) a supervised classifier module, based on K-Means algorithm, to profile network traffic and notify an SLA breach situation to the optimizer module; (ii) a Regressor module which is queried by the optimizer module to predict KPIs of a candidate network topology; and (iii) an optimizer module which uses the Simulated Annealing algorithm to identify the best network topology to comply with the agreed SLA.

Calabrese *et al.* [5] leveraged ML techniques to design a general-purpose learning framework, having the ability to autonomously generate algorithms specialized for Radio Resource Management (RRM) functionalities in 5G RANs. The framework is based on Reinforcement Learning (RL) approach with a decoupling of learning and acting roles of an RL agent. Indeed, the framework architecture consists of one centralized learner and a set of distributed actors. The learner uses experiences sent from actors to learn RRM algorithms, while actors run the RRM algorithms supplied by the learner and repeatedly generate experiences. The separation of learning and acting roles has the advantage of allowing scalability without sacrificing training stability, providing fault-tolerance and enabling transfer learning. To cope with the large dimension of RRM problems, Q-learning via functional approximation of the Q-function is adopted. The ANNs are used as functional approximator owing to their generalization capabilities and the existence of computationally efficient training algorithms. To train ANN, Neural-Fitted Q-iteration (NFQ) is used. Transfer learning, in terms of parameter transfer and instance transfer, is also enabled among actors in the network.

Authors in [6] proposed an anomaly detection and diagnosis solution for holistic RANs self-healing in 5G networks. The anomaly detection is carried out in two steps. First, the profiling of normal system states is performed per cell for work days and weekends. Subsequently, the identification of anomaly patterns is conducted by measuring the deviation from the established baseline profiles. The anomaly diagnosis aims to determine the potential root causes of a detected anomaly. The diagnosis process relies on Case-Based Reasoning (CBR), transfer learning and active learning techniques to allow for autonomous self-healing actions. The holism property, required to build efficient resilient systems, is achieved through cross-domain collaboration. However, a holistic healing cannot be achieved without standardization of management functions and development of mechanisms and KPIs to communicate decisions and actions between management domains.

Qin *et al.* [7] investigated the self-healing problem in SON-based ultra-dense cell networks. ML-based self-healing framework is devised, which provides both outage detection and compensation, even in the presence of partial KPI statistics. The outage detection algorithm applies Support Vector Data Description (SVDD) approach; a ML technique inspired by Support Vector Machine (SVM). The outage compensation is fulfilled through load-balanced allocation of neighboring small-cell resources, guaranteeing coverage and user's QoS requirements.

The work in [8] proposed a reactive mechanism for an adaptive and accurate NFV scaling decisions. The scaling mechanism combines Q-learning and Gaussian process models. As

a reactive solution, it suffers from latency to react to dynamic changes and delay to have new Virtualized Network Function (VNF) instance ready for use. To avoid this weakness, Alawe *et al.* [9] proposed a proactive mechanism based on traffic prediction to enable dynamic scaling of 5G Core Network (CN) resources, particularly Access and Mobility Management (AMF) resources. The forecast of upcoming traffic load and the needed number of AMFs is leaned on two neural network techniques, namely Deep Neural Network (DNN) and Long Short-Term Memory Recurrent Networks (LSTM).

B. Relevant Research Projects

SELFNET [1] is a 5G-PPP phase I project that aims to design and implement an intelligent management framework for 5G networks. The project focuses on the intelligent autonomic management of NFV functions in NFV/SDN-enabled 5G networks. Different use cases have been defined targeting the following capabilities: (i) self-protection against distributed cyber-attacks; (ii) self-healing against network failures; and (iii) self-optimization to dynamically improve the network's performance and the user's QoE. The project defined also a set of Health of Network (HoN) metrics that serve as its KPIs to measure the stability and performance of the network. CogNet [2] is a 5G-PPP phase I project that uses ML to enable self-administration and self-management of 5G networks. The project identified six use cases and eleven scenarios based on the challenges of the future 5G network management, such as network resource utilization, network performance degradation, and energy efficiency. Just-in-Time services and SLA enforcement are among the identified use cases. SLICENET [3] is a 5G-PPP phase II project that aims to design and implement an E2E cognitive vertical-oriented 5G network slicing framework. The project focuses on cognitive network management, control and orchestration of slices across multiple management domains. Cognition (intelligence) is used to learn the best actions to be taken in order to maintain the desired QoS/QoE of the verticals. Three use cases representing three vertical industries are defined, namely: (i) 5G Smart Grid self-healing; (ii) 5G e-Health connected ambulance; and (iii) 5G smart city smart lighting. The project defines the cognition requirements for the aforementioned use cases.

C. Relevant Standards

Along side research contributions, Standards Developing Organisations (SDOs) are progressing standards initiatives looking at autonomous and automated network and service management.

TMF's Zero-touch Orchestration, Operations and Management (ZOOM) project [4] intends to define a new management architecture of virtualized networks and services, based on smooth interaction between physical and virtual components to dynamically assemble into personalized services. ZOOM and ZSM are mainly guided by the same principles, such as dynamic and open APIs, closed-loop end-to-end management, near real-time and zero-touch.

MEF 3.0 [5] is a transformational framework for defining, delivering, and certifying agile, assured, and orchestrated communication services across a global ecosystem of automated networks. MEF 3.0 services will be delivered over automated, virtualized, and interconneceted networks powered by Lifecycle Services Orchestration (LSO), SDN, and NFV. LSO provides open and interoperable automation of management operations for connectivity services. This includes fulfillement, performance, control, assurance, usage, analytics, security, and policy capabilities. MEF is partnering with The Linux Foundation to advance its LSO analytics capabilities using Platform for Network Data Analytics (PNDA) [6].

ETSI ENI (Experiential Network Intelligence) ISG (Industry Specification Group) [7] is defining a Cognitive Network Management architecture using closed-loop AI mechanisms based on context-aware and metadata-driven policies to improve the operator experience. The architecture is based on the "observe-orient-decide-act" control loop model. The project defines different use cases that cover infrastructure management, network operations, service orchestration and management, and assurance. Among ENI use cases that are relevant to ZSM, we find: (i) Intelligent network slicing management; (ii) Network fault identification and prediction; and (iii) Assurance of Tight service requirements. The cognition requirements are identified for different scenarios of service provision and network operation, as well as enabling dynamic autonomous behavior and adaptive policy-driven operation in a changing context. Unlike ETSI ISG ZSM which focuses on automation techniques, end-to-end service management and full automation, ETSI ENI ISG concentrates on AI techniques, policy management and closed-loop mechanisms. The capabilities offered by ENI such as the AI/ML algorithms, intent policies, SLA management can be leveraged by ZSM's analytics and intelligence services to improve the automation of network and service management.

TM Forum Smart BPM [8] (Business Process Management) is investigating the embodying of AI-based decision modeling in telecom business processes such as resource provisioning, fault management, assurance, and customer management. The resulting best practices, tools and methodologies can be applicable to ZSM.

FG-ML5G [9] is an ITU-T Focus Group on Machine Learning for Future Networks that was established in Nov. 2017. The group focuses on defining uses cases and specifying network architectures, interfaces and data formats for enabling machine learning mechanisms in future networks, including 5G. The outputs of FG-ML5G can be exploited by ZSM to enhance the framework intelligence.

[1] https://selfnet-5g.eu
[2] http://www.cognet.5g-ppp.eu
[3] https://slicenet.eu
[4] https://www.tmforum.org/collaboration/zoom-project/
[5] http://www.mef.net/mef30/overview
[6] http://pnda.io
[7] https://www.etsi.org/technologies-clusters/technologies/experiential-networked-intelligence
[8] https://www.tmforum.org/catalysts/smart-bpm/
[9] https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx

III. ZSM Architecture

The ZSM framework reference architecture [10] is designed to support full automated network and service management in multi-domain environments that includes operation across legal operational boundaries. To meet this goal, the design of the ZSM architecture is guided by a set of architectural design principles. The architecture is modular (made up of self-contained and loosely-coupled services), extensible (allowing to add new services and service capabilities), scalable (enabling independent deployment and scaling of components to accommodate the management load) and resilient to failures (where services are designed in a way to cope with the degradation of the infrastructure and/or other services). It is worth noting that modularity is a keystone for achieving the architecture extensibility, scalability and resiliency. The modular characteristic is paired with the use of intent-based interfaces, closed-loop operation and AI/ML techniques to empower full-automation of the management operations. As depicted in Fig. 1, the framework architecture is composed of a set of architectural building blocks, namely, management domains (MDs) including E2E service MD, management services, integration fabric and common data services.

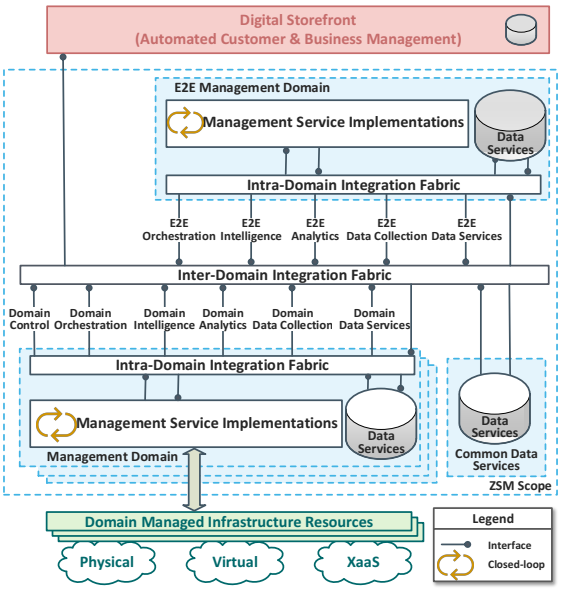


Fig. 1: ZSM Reference Architecture [10].

The ZSM architecture is split into MDs to support the separation of management concerns. Each MD is responsible for intelligent automation of orchestration, control and assurance of resources and services within its scope. The managed resources can be physical resources (e.g., physical network functions (PNFs)), virtual resources (e.g., VNFs) and/or cloud resources (e.g., "X-as-a-service" resources). The E2E service MD is a special MD that manages end-to-end, customer-facing services that span multiple domains provided by different administrative entities. The E2E service coordinates between domains using orchestration. The decoupling of MDs from the E2E service MD escapes monolithic systems, reduces the overall system's complexity, and enables independent evolution of domains and end-to-end management operations.

The common data services allow to separate data storage and data processing, facilitating access to data and cross-domain data exposure. Data in Common Data Services can be exploited by domain and E2E service intelligence services to drive domain-level and cross-domain AI-based closed-loop automation, respectively. The automated decision-making mechanisms are controlled by rules and policies. Note that a MD may contain domain data services that allow data sharing between functional components inside the MD.

Each MD, including the E2E service MD, comprises several management functions grouped into logical groups (e.g., domain collection services, domain analytics services, domain intelligence services, domain orchestration services and domain control services) and provides a set of management services via service interfaces. Some services are only provided and consumed locally inside the domain using the intra-domain integration fabric. Meanwhile, the service exposure cross-domain is enabled through inter-domain integration fabric. The management services are exposed and consumed following either the request-response or the publish-subscribe patterns. Fig. 2 illustrates the high-level architecture of a MD with the main interactions between the different logical groups of management services.

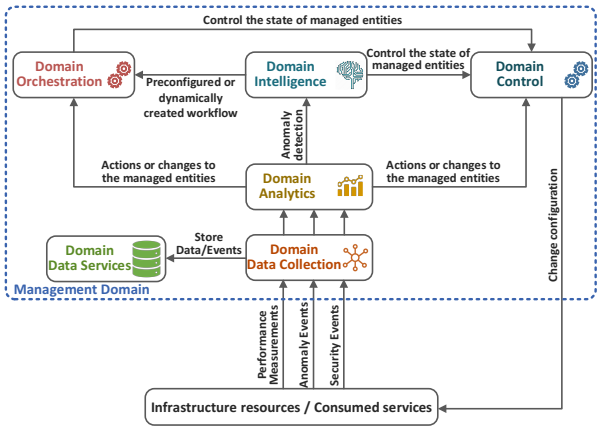


Fig. 2: Management Domain's High-Level Architecture.

IV. Limitations and Risks of AI-driven ZSM

AI plays a pivotal role in empowering self-managing functionalities in ZSM, leading to improved service delivery and reduced OPEX. However, leveraging AI techniques in a ZSM system is constrained by several limitations and risks, highlighted in what follows.

A. Limited Automation due to Limited AI

AI/ML mechanisms are tremendous for embodying cognitive processing to ZSM systems and allowing full automation of management operations. However, this goal can not be met without addressing limitations associated with AI/ML techniques to fulfill performance and legal requirements. In fact, network operators call for high level of reliability and service availability to avoid financial losses due to network outages and SLA violations. Moreover, transparency and accountability of AI-enabled systems are of utmost importance

to build trust in their decisions as well as legal compliance. For instance, the General Data Protection Regulation (GDPR) law entitles individuals the right to obtain an explanation of how the decision is reached by an automated system. In what follows, we will discuss different limitations of AI/ML models that can hamper the fulfillment of the aforementioned requirements.

1) Lack of Datasets and Labeling: The validation and accuracy of ML models heavily depend on the availability of high-quality datasets. Thus, 5G-specific datasets are crucial for building up efficient and accurate learning models in a ZSM system. Unfortunately, such datasets are actually scarce since the roll-out of 5G networks is planned for 2020. Moreover, the existing operators' data are not accessible due to privacy issues. Even some recent initiatives (e.g., 5GMdata [10] and 5GTN [11]) are raising to create 5G-specific datasets, the generated samples are synthetic and/or lack completeness. Besides data availability, the quality of collected data is another issue. Indeed, high-quality data; i.e., accurate, suitable, complete and timely, is necessary for delivering useful insights and decisions. Another challenge lies in the volume of data needed to reach high accuracy. For instance, the deep learning accuracy scales with the amount of available data; the higher the volume of trained data, the higher the accuracy will be. The supervised and semi-supervised learning adds another layer of complexity as labeled data is required to train algorithms. Annotated data may be scarce or expensive, and a fully annotated dataset may not be feasible.

2) AI Model Interpretability: The adoption of AI/ML techniques to enable full automation in ZSM will potentially depend on how well AI/ML models are interpretable. The AI/ML model interpretability is the ability to establish the cause-and-effect relationship between decisions made and input data that caused such decisions. It is the process of explaining *what*, *how* and *why* decisions are taken based on training data. The AI/ML model interpretation will ensure accountability, reliability and transparency, fostering trustworthiness in AI-enabled systems. Unfortunately, the interpretation of an AI/ML model is a challenging task that can not be achieved without sacrificing the model accuracy. Indeed, simple models like linear and tree-based models are easily interpretable but suffer from low accuracy. For instance, the response function (i.e., predicted output) produced by a linear model is expressed as a weighted sum of its input data (i.e., features), which makes the model interpretability a straightforward process. However, linear models fail to capture complex non-linear patterns in the data, leading to drop in accuracy. Meanwhile, more complex models such as ensemble and DL models often yield higher accuracy, thanks to their ability to learn complex non-linear relationships between inputs and outputs. Neverthless, those models are reputed to be black-box models as the logic behind their decisions is extremely difficult to explain. Taking the example of DL models, it is quite hard to understand the role played by individual neurons and the correlation between input features and model decisions. Therefore, a tradeoff between interpretability and accuracy should be established.

[10]https://github.com/lasseufpa/5gm-data/wiki/5GMdata-Home

[11]https://5gtn.fi

3) Training Time and Inference Accuracy: To support the promised very-low latency and ultra-low reliability of next-generation networks, a ZSM system should enable real-time or near real-time management operations with accurate decision making. Emerging AI/ML techniques, such as ensemble and deep learning, have proved their capability in solving complex real-world problems with high accuracy. Therefore, those techniques are likely to be a key enabler in a ZSM system to deliver accurate decisions. However, the lengthy training time required to achieve such improved accuracy may jeopardize their practicality for real-time usage. This issue becomes even more critical in a highly dynamic and non-stationary environment, such as 5G networks. In such environment, the data patterns may change over time, calling for AI/ML model retraining to accommodate the new changes in data distribution and consequently achieve higher prediction (or inference) accuracy. While model retraining prevents drop in AI/ML model performance, it entails a considerable increase in training time. Thus, shortening the training time without loss of inference accuracy is essential for emerging AI/ML techniques to make their way into a ZSM system.

4) Computation Complexity: Emerging AI/ML techniques, such as deep learning and reinforcement learning, have gained a surge of interests thanks to their noticeable accuracy improvements. However, this accuracy enhancement comes at the cost of high demand of computation, memory and energy resources. Thus, leveraging such techniques by a ZSM system is challenged by the near-to-zero latency and lower energy usage promises of 5G networks. To fully benefit from those models, efficient solutions to optimize and accelerate them are crucial.

B. Security

As shown in the aforementioned projects and contributions, AI/ML techniques play an important role in empowering functionalities such as self-planning, self-optimization, self-healing, and self-protecting. However, the growing enthusiasm for AI/ML adoption in managing next-generation networks could be waned if security concerns related to the use of AI/ML techniques are not addressed. Indeed, the use of AI/ML and other data analytic technologies is a source for new attack vectors in a ZSM system [11]. It has been proven that ML techniques are vulnerable to several attacks [12] targeting both training phase (i.e., poisoning attacks) and test phase (i.e., evasion attacks). The attacks aim to cause either integrity, availability or privacy violation by introducing carefully crafted perturbations to training and test samples. Such perturbations are called adversarial examples.

The ZSM's E2E service intelligence services drive the closed loops in the E2E service management domain. It covers both service-specific predictions / recommendations (e.g., predict service demand) and making decisions and triggering their execution (e.g., decisions to optimize the E2E service) [10]. The decision-making is based on information obtained from domain data collection services and common data services. Thus, an attacker can craft inputs to drive the ML model used by the E2E service intelligence services to make erroneous prediction and decision-making, potentially

causing performance degradation and financial harm, as well as endangering SLA fulfillment and security guarantees. For instance, an adversary can inject crafted samples that let the E2E service intelligence services wrongly forecast the future resource requirements of an E2E service, or to trigger an inappropriate management policy (e.g., reconfiguration, scale-in, scale-out) of E2E service. Note that domain intelligence services are prone to the same attacks. Fig. 3 illustrates a VNF auto-scaling scenario where the ML model, used by the E2E service intelligence or domain intelligence, generates scaling decisions (i.e., scale-in, scale-out) based on service requirements and VNF load. As shown in the figure, if metric data are not manipulated, the ML model will decide to perform a scale-in operation to reduce costs. However, if an attacker is able to manipulate the metric data, he/she could fool the ML model into taking the scale-out decision, which will result in adding new VNF instances. To make things worse, the attacker may leverage the closed-loop feature to generate and inject adversarial inputs automatically and repeatedly into the system, which may result in Denial of Service (DoS) due to resource exhaustion and/or increased OPEX.

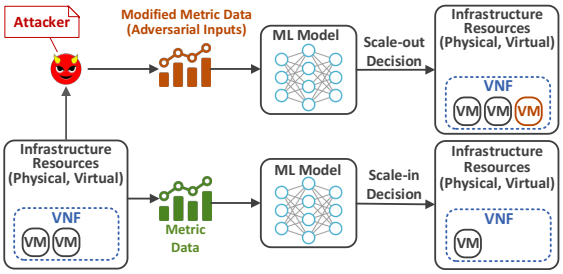


Fig. 3: Adversarial Attack Illustrative Example.

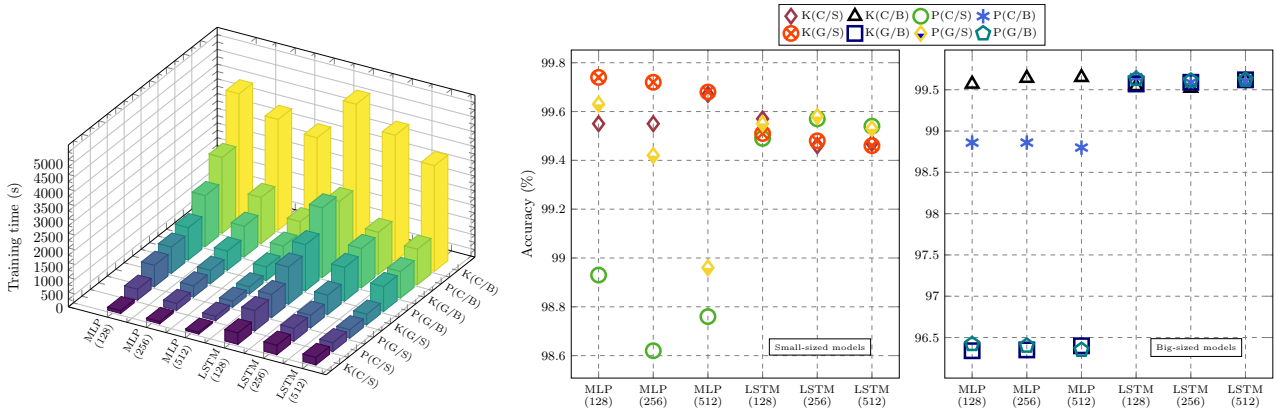
V. Impact of Hardware/Software Factors on Training Time and Inference Accuracy: A Benchmark Study

Many factors may contribute to speeding up the training time, including the dataset size, the development platform (e.g., Tensorflow, Pytorch, Keras, Caffe, MXNet), the hardware platform (e.g., CPU, GPU, TPU) and the AI/ML model's hyper-parameters (e.g., the number of hidden layers, the number of neurons in each layer, the learning rate, the batch size, the number of epochs). In this section, a benchmark study is conducted to explore how some of the aforementioned factors influence the training time and inference accuracy of a DL-based DoS detection model. The recent intrusion detection evaluation dataset, CICIDS2017 [13] [12], is used. CICIDS2017 dataset comprises benign traffic and the most up-to-date common attacks. For the purpose of this study, we consider a subset of CICIDS2017 dataset, where only network flows corresponding to normal traffic and DoS/Distributed DoS (DDoS) attacks are kept. The dataset is prepared to fit for DL models by performing different data processing operations, namely: removing records that are redundant or have missing/infinity values, encoding non-numerical features, and normalizing the values of features using the Min-Max scaling technique. The resulting dataset contains a total of 961641 flows, where each flow is defined by a feature vector containing 79 features in addition to a label identifying the flow's class (i.e., benign or malicious). The dataset is split into separate training and test sets with a ratio of $0.7/0.3$, respectively. The DL model is built on the training set while its prediction accuracy is evaluated on the unseen test dataset. To investigate the effect of the type and architecture of the chosen DL algorithm, four model variants are considered: (1) *Small MLP* (MultiLayer Perceptron), involving 1 input layer and 2 hidden layers with 64 neurons each; (2) *Big MLP*, consisting of 1 input layer and 4 hidden layers with 1024 neurons each; (3) *Small LSTM*, comprised of 2 hidden LSTM layers with 120 cells each; and (4) *Big LSTM*, composed of 4 hidden LSTM layers with 384 cells each. The four variants used a two-class softmax output layer. The models are implemented using the Python's DL libraries Pytorch and Keras running on a TensorFlow backend. They are trained for 10 epochs, with different batch sizes (128, 256, and 512). The experiments are carried out on two platforms, namely: (1) a VM with 16-cores Intel's Skylake $2.4GHz$ CPU and $64GB$ RAM, and (2) a NVIDIA Jetson TX2 GPU with 256 CUDA cores, $8GB$ RAM and JetPack $4.2.1$ SDK.

Figures 4a and 4b depict the comparative results on training speed and inference accuracy, respectively. The analysis of the obtained results has revealed the following key insights: (1) The training time can be shortened by reducing the model size and increasing the batch size; (2) CPU outperforms GPU in speeding-up the training time of small-sized models. This stems from the fact that with small models, the CPU-GPU data transfer overhead exceeds the computation acceleration benefit; (3) LSTM-based models exhibit long training time compared to MLP-based models; (4) Pytorch performs well, in terms of training speed, for big-sized models as well as when running on GPU. Indeed, Pytorch-based big models trained on GPU could achieve up to $7.5\times$ (LSTM) and $4.3\times$ (MLP) speedup compared to their Keras-based counterparts executed on CPU. However, up to 3.4% loss of accuracy is observed with Pytorch-based MLP models. The effect of varying the training set size on the training speed and prediction accuracy is also assessed. Figure 5 shows the results for a *Big MLP* model using a batch size of 512 and running on GPU for different training set sizes ranging from 10000 to 670744. Typically, the availability of more training data leads to improved accuracy, but at the expense of increased training time. A key observation is that, unlike Keras, Pytorch-based implementation delivers faster training speed and its accuracy is less sensitive to the amount of training data.

In the light of this study, it is clear that finding the ideal combination of the diverse software/hardware factors is an essential, yet a challenging, task to achieve the required accuracy at fast training speed. Considering further performance metrics, such as inference time or energy consumption, makes this task even more difficult.

[12]https://www.unb.ca/cic/datasets/ids-2017.html



(a) (b)

Fig. 4: Comparative results on training speed and inference accuracy of a DL-based DoS detection model (P=Pytorch, K=Keras, C=CPU, G=GPU, S=Small and B=Big).

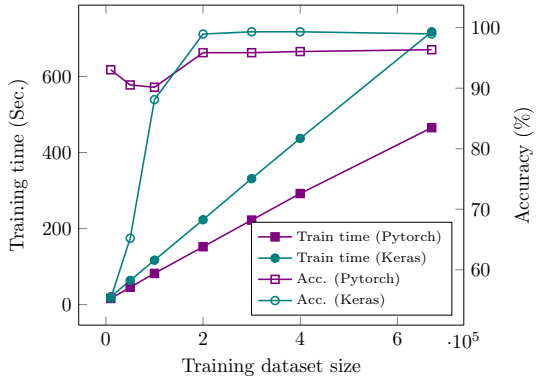


Fig. 5: Speed/Accuracy vs. training set size of Big MLP model with a batch size of 512 on GPU.

VI. Future Research Directions

A. Safe Shared Learning

Collaboration and data sharing between multiple mobile operators (i.e., different MDs) are vital to improve accuracy and speed up the learning process of ML models used by the different MDs. Meanwhile, empowering shared learning gives rise to privacy and trust issues. It is necessary to ensure that the model can learn from shared data without compromising the privacy of collaborative entities. To deal with the trust issue, mechanisms to ensure that collaborative entities are not malicious need to be developed.

B. Trust in Data and Models

Trust is a cornerstone for adopting AI-based automated systems. Two dimensions of trust are necessary to stimulate confidence in AI-based systems, namely trust in datasets and trust in AI models.

The efficiency of predictions and decisions made by ZSM's intelligence services will depend on data gathered from a variety of sources (e.g., users, services, network) across multiple domains. Since data is the fuel for AI algorithms, it is crucial to ensure their integrity and their provenance from trusted sources. Thus, solutions to automatically collect trusted immutable datasets from distributed sources are necessary. Blockchain can play a pivotal role in developing such solutions thanks to its immutability and distributed nature.

The trust in AI models is related to which extent domain experts can trust the decisions taken by those models. As mentioned before, the model interpreatbility is a key enabler to foster trust in AI-enabled systems. However, the complexity of emerging ML techniques, such as deep learning and reinforcement learning, is a serious impediment to their interpretability. They are considered black-boxes since it is hard to explain how they work and how their outcomes are made. To fully benefit from the high accuracy brought by black-box models in a ZSM system, it is necessary to design efficient interpretation approaches that improve explanation of black-box models without sacrificing their accuracy. It is desirable to be able to generate interpretations automatically.

C. Computation Complexity Optimization

As already mentioned, the emerging ML techniques are characterized by increased accuracy but at the cost of high demand of computation resources. To make their adoption possible in a ZSM system, solutions to optimize and accelerate their execution are necessary. Thus, new optimization techniques should be designed to reduce the complexity of those models without loss of accuracy. A possible optimization consists in reducing the number of operations required by the model. Hardware-based methods (e.g., FPGA-based acceleration and GPU processing) to accelerate complex ML models is another alternative to explore.

D. Training Speed – Accuracy Balance

To empower ZSM's analytics and intelligence services that can take advantage of emerging AI/ML techniques while meeting both (near) real-time and accurate prediction/decision making requirements, a balance should be established between the training time and the accuracy of the integrated AI/ML

models. Along with optimizing the computation complexity as discussed above, a potential paradigm that has recently surfaced as a promising solution to tackle the slow training issue is transfer learning.

Transfer learning consists in leveraging the knowledge acquired from one task to solve a new but related task. For instance, the experience gathered by an AI/ML model trained to detect DoS attacks in CN can be shared with a newly deployed AI/ML model aiming to detect DoS attacks in RAN. Using a pre-trained AI/ML model to predict new data patterns (e.g., new classes of attacks) is yet another possible application of transfer learning. The capability of transferring previous knowledge leads to fast training process and improved accuracy of the new model. However, a major challenge for applying transfer learning paradigm is to identify *what*, *how* and *when* to transfer knowledge in order to avoid a negative effect on the performance of the new model. Thus, more research efforts in this direction are required to fulfill the potential of transfer learning in a ZSM system.

E. Adversarial ML for ZSM

To make ML techniques resilient to adversarial attacks, a new research discipline has emerged, called Adversarial Machine Learning (AML) [14]. It aims at assessing the security robustness of ML algorithms against attacks and designing appropriate countermeasures. While AML has attracted much interest in vision field, only very few contributions (e.g., [15], [16]) have addressed ML security in the context of service and network management. Usama et al. [15] highlight the importance of tackling adversarial attacks against cognitive self-organizing networks. As a proof of concept, white-box evasion attacks against Convolutional Neural Network have been designed to show how a malware classifier can be evaded. Han *et al.* [16] investigated the reaction of Reinforcement Learning (RL) agent toward different forms of causative attacks in the context of autonomous cyber-defense in Software Defined Networks (SDNs). Guaranteeing the security of ML models is a mandatory condition for their integration in a service and network management platform for next-generation networks. Thus, more research efforts in AML need to be devoted to this area. Indeed, we need to master how adversarial attacks can be launched in networking environment. While the generation of adversarial examples is now relatively clear in vision area, there are no clues on how they could be crafted and introduced in a network traffic. Research work should focus on devising algorithms that automatically generate adversarial examples for network traffic. Moreover, suitable countermeasures should be designed to cope with those attacks and ensure the safety of ML models integrated in a ZSM system. Another research direction is to propose a certification framework to assess the security properties of ML techniques.

F. Learning Correctly in the Presence of Adversaries

Expecting that we will be able to get rid of all possible adversarial attacks against ML models by elaborating strong countermeasures is unrealistic. Hence, an important question arises on how we can learn correctly in the presence of adversarial examples. In other words, how can we make learning more robust/secure so that the system can take the right decision even in the presence of adversaries?

VII. Conclusion

This paper introduced the emerging concept of Zero-touch network and Service Management (ZSM). We showed that AI techniques play a pivotal role in making ZSM a reality. Meanwhile, we spotlighted the limitations and security risks that may hamper the integration of AI techniques in ZSM. In light of the identified issues, we pointed out potential research topics. A special attention should be paid to devise computationally efficient and trustable AI-driven network management operations.

Acknowledgment

This work was supported in part by the Academy of Finland Project 6Genesis Flagship (Grant No. 318927), CSN (Grant No. 311654) and the European Union's Horizon 2020 research and innovation programme under the INSPIRE-5Gplus project (Grant No. 871808).

References

[1] 5G-PPP Architecture WG, "View on 5G Architecture (Version 2.0)," July 2017.

[2] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2432 – 2455, 4Q 2017.

[3] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization," in *INFOCOM*. IEEE, 2017, pp. 1–9.

[4] A. Martin, J. Egaña, J. Flórez, J. Montalbán, I. G. Olaizola, M. Quartulli, R. Viola, and M. Zorrilla, "Network Resource Allocation System for QoE-Aware Delivery of Media Services in 5G Networks," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 561 – 574, June 2018.

[5] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning Radio Resource Management in RANs: Framework, Opportunities, and Challenges," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 138 – 145, Sept. 2018.

[6] J. Ali-Tolppa, S. Kocsis, B. Schultz, L. Bodrog, and M. Kajo, "Self-healing and Resilience in Future 5G Cognitive Autonomous Networks," in *10th ITU Academic Conf., Machine Learning for a 5G Future*, Nov. 2018, pp. 35 – 42.

[7] M. Qin, Q. Yang, N. Cheng, H. Zhou, R. R. Rao, and X. Shen, "Machine Learning Aided Context-Aware Self-Healing Management for Ultra Dense Networks with QoS Provisions," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12 339 – 12 351, Dec. 2018.

[8] C. H. T. Arteaga, F. Rissoi, and O. M. C. Rendon, "An Adaptive Scaling Mechanism for Managing Performance Variations in Network Functions Virtualization: A Case Study in an NFV-Based EPC," in *Proc. of the IEEE 13th Int. Conf. on Network and Service Management*, 2017, pp. 1 – 7.

[9] I. Alawe, A. Ksentini, Y. Jadjadj-Aoul, and P. Bertin, "Improving traffic forecasting for 5g core network scalability: A machine learning approach," *IEEE Network Magazine*, pp. 1 – 10, 2018.

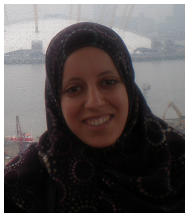
[10] ETSI GS ZSM 002, "Zero-touch Network and Service Management (ZSM); Reference Architecture," Aug. 2019.

[11] C. Benzaid, and T. Taleb, "ZSM Security: Threat Surface and Best Practices," *IEEE Network Magazine*, to appear.

[12] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. Tygar, "Can Machine Learning Be Secure?" in *Proc. of ASIACCS'06*, 2006, pp. 16 – 25.

[13] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *In Proc. of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, Jan. 2018.

[14] L. Haung, A. D. Joseph, B. Nelson, B. I. Rubinstrein, and J. D. Tygar, "Adversarial Machine Leaning," in *In Proc. of 4th ACM Workshop on Artificial Intelligence and Security,*, Oct. 2011, pp. 43 – 58.
[15] M. Usama, J. Qadir, and A. Al-Fuqaha, "Adversarial attacks on cognitive self-organizing networks: The challenge and the way forward," *CoRR*, vol. abs/1810.07242, 2018.
[16] Y. Han, B. I. Rubinstein, T. Abraham, T. Alpcan, O. De Vel, S. Erfani, D. Hubczenko, C. Leckie, and P. Montague, "Reinforcement Learning for Autonomous Defence in Software-Defined Networking," in *In Proc. of the 9th Int. Conf. on Decision and Game Theory for Security (GameSec)*, Aug. 2018, pp. 145 – 165.



Chafika Benzaïd is currently a PostDoc researcher at MOSA!C Lab, Aalto University. She is an associate professor and research fellow in Computer Science Department at University of Sciences and Technology Houari Boumediene (USTHB). She obtained her PhD degree in Computer Sciences from USTHB in 2009. Her current research interests include AI-driven network security and AI security. She serves/served as a TPC member for several international conferences and as a reviewer for multiple international journals.

Tarik Taleb is Professor at Aalto University and University of Oulu. He is the founder and director of the MOSA!C Lab (www.mosaic-lab.org). Prior to that, he was a senior researcher and 3GPP standards expert at NEC Europe Ltd., Germany. He also worked as assistant professor at Tohoku University, Japan. He received his B.E. degree in information engineering, and his M.Sc. and Ph.D. degrees in information sciences from Tohoku University in 2001, 2003, and 2005, respectively.