The Hexa-X project vision on Artificial Intelligence and Machine Learning-driven Communication and Computation co-design for 6G

Mattia Merluzzi, Tamas Borsos, Nandana Rajatheva, Andras Benczur, Hamed Farhadi, Taha Yassine, Markus Dominik Mueck, Sokratis Barmpounakis, Emilio Calvanese Strinati, Dilin Dampahalage, et al.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Pre-print submitted to IEEE Access

Digital Object Identifier 10.1109/ACCESS.2022.0092316

The Hexa-X project vision on Artificial Intelligence and Machine Learning-driven Communication and Computation co-design for 6G

MATTIA MERLUZZI[1], (Member, IEEE), Tamás Borsos[2], Nandana Rajatheva[3], (Senior Member, IEEE), András A. Benczúr[4] , Hamed Farhadi[5], Taha Yassine[6,7], Markus Dominik Mück[8], (Member, IEEE), Sokratis Barmpounakis[9], Emilio Calvanese Strinati[1], (Member, IEEE), Dilin Dampahalage[3], (Graduate Student Member, IEEE), Panagiotis Demestichas[9], (Senior Member, IEEE), Pietro Ducange[10], Miltiadis C. Filippou[8], (Senior Member, IEEE), Leonardo Gomes Baltar[8], (Senior Member, IEEE), Johan Haraldson[5], Leyli Karaçay[11], Dani Korpi[12], (Member, IEEE), Vasiliki Lamprousi[9], Francesco Marcelloni[10], (Member, IEEE), Jafar Mohammadi[13], (Member, IEEE), Nuwanthika Rajapaksha[3], (Graduate Student Member, IEEE), Alessandro Renda[10], Mikko A. Uusitalo[12], (Senior Member, IEEE)

[1]Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France (e-mail: mattia.merluzzi@cea.fr)
[2]Ericsson Research, Hungary
[3]Centre for Wireless Communications, University of Oulu, Finland
[4]ELKH Institute for Computer Science and Control, Hungary
[5]Ericsson Research, Stockholm, Sweden
[6]Institute of Research & Technology b<>com, Cesson-Sévigné, France
[7]Univ Rennes, INSA Rennes, CNRS, IETR –UMR 6164, F-35000 Rennes, France
[8]Intel Deutschland GmbH, Munich, Germany
[9]WINGS ICT Solutions, Athens, Greece
[10]Department of Information Engineering, University of Pisa, Italy
[11]Ericsson Research, Turkey
[12]Nokia Bell Labs, Espoo, Finland
[13]Nokia Bell Labs, Stuttgart, Germany

Corresponding author: Mattia Merluzzi (e-mail: mattia.merluzzi@cea.fr).

This work has been partly funded by the European Commission through the H2020 project Hexa-X (Grant Agreement no. 101015956). Miltiadis C. Filippou is now with Nokia Bell Labs, Munich, Germany.

ABSTRACT This paper provides an overview of the most recent advancements and outcomes of the European 6G flagship project Hexa-X, on the topic of in-network Artificial Intelligence (AI) and Machine Learning (ML). We first present a general introduction to the project and its ambitions in terms of use cases (UCs), key performance indicators (KPIs), and key value indicators (KVIs). Then, we identify the key challenges to realize, implement, and enable the native integration of AI and ML in 6G, both as a means for designing flexible, low-complexity, and reconfigurable networks (*learning to communicate*), and as an intrinsic in-network intelligence feature (*communicating to learn* or, 6G as an efficient AI/ML platform). We present a high level description of down selected technical enablers and their implications on the Hexa-X identified UCs, KPIs and KVIs. Our solutions cover lower layer aspects, including channel estimation, transceiver design, power amplifier and distributed MIMO related challenges, and higher layer aspects, including AI/ML workload management and orchestration, as well as distributed AI. The latter entails Federated Learning and explainability as means for privacy preserving and trustworthy AI. To bridge the gap between the technical enablers and the 6G targets, some representative numerical results accompany the high level description. Overall, the methodology of the paper starts from the UCs and KPIs/KVIs, to then focus on the proposed technical solutions able to realize them. Finally, a brief discussion of the ongoing regulation activities related to AI is presented, to close our vision towards an AI and ML-driven communication and computation co-design for 6G.

INDEX TERMS connecting intelligence, 6G networks, sustainability, trustworthiness, energy efficiency, AI and ML for air interface design, edge AI, explainable AI

I. INTRODUCTION

Today, we are at the early stage of the research on the sixth generation of mobile networks (6G), whose standardization activities (which however have not yet kicked off) are expected to deliver first specifications in 2030 [1]. Despite the long-term timeline, several players and stakeholders from academia [2]–[4] to industry [5]–[10], including collaborative projects through public initiatives and funding[1], have already started cogitating on new enabled services and Use Cases (UCs), along with the respective key Performance Indicators (KPIs) and Key Value Indicators (KVIs), with the latter being novel measures of how future wireless networks can help addressing various societal needs (values) such as sustainability, trustworthiness, flexibility, and inclusion [11].

In this context, the European 6G Flagship project Hexa-X [11] investigates several features of future communication systems, and its general goal is to harmonize the global 6G vision to define an intelligent fabric of technology enablers connecting human, physical, and digital worlds[2], with values comprising sustainability, trustworthiness, and inclusion. Hexa-X covers, across seven work packages: *i)* UCs (organized in Use Case Families - UCFs), KPIs, KVIs and general architectures *ii)* radio access technologies, *iii)* localization and sensing, iv) in-network Artificial Intelligence (AI) and Machine Learning (ML) *v)* architectural aspects, *vi)* orchestration and management, and *vii)* special purpose functionalities.

Beyond the project ecosystem, in the overall research landscape, while all contributions and technical proposals obviously differ in several aspects, none of them disagrees (despite slightly different terminologies) on the fact that ML and AI will be indispensable and native components of 6G, towards a new paradigm shift, from the legacy concept of connecting humans and things, to the new challenge of *connecting intelligence* [12]. As already mentioned, to recognize the fundamental role of ML and AI in 6G, Hexa-X dedicates an entire work package (i.e., Work Package 4, entitled *AI-driven communication and computation co-design*), to AI and ML in 6G, especially from an algorithmic perspective [13], [14]. Part of its outcomes and harmonized view are the focus of this paper.

Besides the technical challenges, the integration of AI and ML into wireless networks needs a concrete definition of novel KPIs, KVIs and, in general, metrics able to properly assess the performance (and values) of AI and ML-based methods in 6G. Also, *learning*, can play a twofold role in 6G [5], [12], [15]–[17]: *i) Learning to Communicate* (L2C), which is about applying AI and ML-based methods to enhance network performance with extreme flexibility and low complexity, and *ii) Communicating to Learn* (C2L), which is about conceiving 6G networks as enablers of AI and ML-based services. For both paradigms, a rethinking of the network is needed at all levels: from the physical layer, with new channel estimation techniques, Power Amplifier (PA) non-linearity compensation, advanced beamforming in (massive) Multiple-Input-Multiple-Output (MIMO) settings, etc., up to the highest layers, including the ever tighter integration of computing and storage capabilities into communication networks. The latter will be (jointly) optimized and orchestrated with wireless resources, to achieve new challenging targets of performance (also in terms of communication task effectiveness), energy efficiency, sustainability, trustworthiness, privacy, and security.

A. STATE OF THE ART

The integration of AI and ML in wireless networks is not a new research topic, and has raised considerable interest in the last few years from several perspectives, including architectural and algorithmic ones, UCs, and standardization [12]. [4] provides an overview of the main technological pillars of 6G, among which AI and ML are considered as enablers of semantic communications [18] and self-organizing networks. In [19], an overview of the main research trends related to 6G is provided, also covering AI and ML as technological enablers, while the focus of [20] is strictly related to AI. Ten challenges related to the integration of ML in 6G are also presented in [17]. While several works cover the L2C concept, the C2L paradigm and the challenges associated with edge AI are also widely discussed [12], [15], [16]. In [16], edge computing is presented as a key enabler of edge intelligence with different possible tiers, answering questions on where to place and run learning and inference tasks (e.g., fully in the cloud, fully at the edge, or hybrid solutions). Task-oriented communication, orchestration aspects, and data governance, are discussed in [21], which also introduces the concept of 6G as an XaaS (i.e., Everything-as-a-Service) platform. Other open challenges in deploying ML in wireless networks, and especially standardization activities, are discussed in [22]. Also, the concept of explainable AI, which is also focus of the present paper, is discussed in [23]. One of the most recent and comprehensive vision papers on 6G can be found in [12], which provides a comprehensive and high level overview of the enablers of both paradigms (i.e., L2C and C2L), with architectures, algorithms, requirements, standardization, platforms, and applications.

While all these works provide general overviews of the envisioned 6G ecosystem and enablers, they lack specific focus on the development of technical solutions to enable identified classes of use cases, along with the associated results showing feasibility and performance.

The goal of this paper is to fill the gap between overview efforts and purely technical contributions, presenting a unified view of the levers at different network layers, along with their integration into an architecture (i.e., the Hexa-X's one) and its view of UCs, KPIs, and KVIs.

B. CONTRIBUTION: THE HEXA-X VISION

Differently from the state of the art, this work represents the common vision of the Hexa-X consortium, specifically that

[1] https://5g-ppp.eu/5g-ppp-phase-3-6-projects/
[2] https://hexa-x.eu/

of the work package committed to the AI and ML-related research from an algorithmic perspective, with a top-down approach that starts with the UCs, and covers the technical solutions with representative numerical results, to show the relevance to the discussed KPIs and KVIs. Our common vision orbits around the following KPIs and KVIs: *throughput, reliability, complexity reduction, accuracy, energy efficiency, privacy, and trustworthiness*. In summary, it includes aspects related to the L2C paradigm with *i)* physical layer, and in particular channel estimation in various scenarios, *ii)* radio transceivers, with air interface design and PA non-linearity compensation, and *iii)* Distributed-MIMO (D-MIMO) settings with resource allocation and beam selection. Then, the C2L vision covers *iv)* workload management, including the AI-as-a-Service (AIaaS) concept, the resilient deployment of distributed AI, the workload placement with energy efficiency targets, load balancing issues in FL settings, and the joint orchestration of radio and computing resources for edge inference; also, it covers *v)* trustworthy and distributed AI, including resilience to adversarial attacks, and (federated) explainable AI. Finally, relevant regulation and standardization aspects are also discussed.

The contribution of this paper is summarized as follows:

- We present a harmonized view of UCs, KPIs, and KVIs related to AI and ML related activities in Hexa-X, which has been developed during the project lifetime.
- We discuss how the developed technical solutions can enable the down selected UCs with the target requirements. This is done by mapping the technical enablers to the Hexa-X architecture, UCs, KPIs, and KVIs, but also the metrics exploited in this paper to assess these indicators, with the latter being *quantified performance* that act as proxies of the proposed KPIs and KVIs.
- Thanks to these metrics, and going beyond our precursor conference paper [24], we show representative evaluation of performance, addressing aspects related to throughput, reliability (e.g., block error rate - BLER), complexity, accuracy, trustworthiness, and energy efficiency.
- We present part of the regulation actions related to AI, focusing on activities of the European Commission (EC) with the so called AI Act [25], and discussing their potential impact to future communication networks.

C. ORGANIZATION OF THE PAPER

The remainder of this paper is organized as follows: Section II presents a high level overview of the UCs, KPIs, and KVIs, with specific yet general references to the AI and ML-related work. It also maps the technical enablers presented in the next sections of this paper to the architecture, UCFs and their UCs, KPIs, and KVIs. Section III is the first technical section and its focus is the L2C paradigm. It provides, after a more detailed overview on the UCs, KPIs and KVIs related to the L2C paradigm, a technical description of the proposed solutions, with the latter spanning across different layers of the protocol stack. Several new, beyond 5G metrics are introduced and associated performance evaluations are presented. Following the approach of Section III, Section IV provides the same kind of analysis of the technical enablers related to the C2L paradigm. From a technical point of view, both sections provide a high level description of the proposed solutions, entailing a brief state of the art, the description of the technical enabler, and numerical results assessing the performance in terms of the identified KPIs and KVIs. To complement the technical sections, Section V presents an overview of the currently ongoing regulation activities related to AI and ML, to discuss their implications on communications related research. Finally, Section VI draws the conclusions and proposes some future directions. Acronyms will be defined the first time they appear in the text, and are also reported in alphabetical order in Table 1.

TABLE 1. List of acronyms

5G	Fifth generation of mobile systems	MEH	Mobile Edge Host
5GAA	5G Automotive Association	MIMO	Multiple-Input-Multiple-Output
6G	Sixth generation of mobile systems	MIP	Mixed Integer Programming
AI	Artificial Intelligence	ML	Machine Learning
AIaaS	AI-as-a-Service	MP	Matching Pursuit
AIS	AI Information Service	MSE	Mean Squared Error
ACLR	Adjacent Channel Leakage Ratio	MMSE	Minimum Mean Squared Error
ANN	Artificial Neural Network	MU-MIMO	Multi_User MIMO
AoA	Angle of Arrivals	MTD	Moving Target Defense
AP	Access Point	NMSE	Normalized Mean Squared Error
API	Application Programming Interface	NN	Neural Network
AR	Augmented Reality	OF	Objective Function
B5G	Beyond 5G	OFDM	Orthogonal Frequency Division Multiplexing
BER	Bit Error Rate	OMP	Orthogonal Matching Pursuit
BLER	Block Error Rate	PA	Power Amplifier
C2L	Communicating to Learn	PAE	Power Added Efficiency
CDF	Cumulative Distribution Function	PET	Privacy Enhancing Technologies
CEE	Channel Estimation Error	QAM	Quadrature Amplitude Modulation
CFO	Carrier Frequency Offset	QoR	Quality of Results
CNN	Convolutional Neural Network	QoS	Quality of Service
CPU	Central Processing Unit	RAN	Radio Access Network
CS	Compressed Sensing	RF	Radio Frequency
CSI	Channel State Information	SCO	Sampling Clock Offset
DT	Digital Twin	SDGs	Sustainable Development Goals
D-MIMO	Distributed-MIMO	SE	Spectral Efficiency
DL	Donwlink	SGD	Stochastic Gradient Descent
DNN	Deep Neural Network	SHAP	Shapley Additive exPlanations
DPD	Digital-pre-distortion	SISO	Single-Input-Single-Output
E2E	end-to-end	SFC	Service Function Chain
EC	European Commission	SNR	Signal-ot-Noise Ratio
ECF	Estimate-compress-forward	SNN	Spiking Neural Network
ESO	European Standardization Organization	SR	Standardization Request
ETSI	European Telecommunications Standards Institute	SRS	Sounding Reference Signal
FedAvg	Federated Averaging	TSK-FRBS	Takagi–Sugeno-Kang Fuzzy Rule-Based Models
Fed-XAI	Federated Explainable AI	THz	TeraHertz
FL	Federated Learning	UC	Use Case
GBT	Gradient Boosted Tree	UCF	Use Case Family
IEEE	Institute of Electrical and Electronics Engineers	UE	User Equipment
ILP	Integer Linear Programming	UL	Uplink
KIP	Key Isolator Partitioner	UN	United Nations
KPI	Key Performance Indicator	V2X	Vehicle-to-Everything
KVI	Key Value Indicator	VNF	Virtual Network Function
LASSO	Least Absolute Shrinkage and Selection Operator	VR	Virtual Reality
LISTA	Learned Iterative Shrinkage and Thresholding Algorithm	VSF	Virtual Security Functions
L2C	Learning to Communicate	WMMSE	Weighted Minimum Mean Squared Error
LDPC	Low-density parity-check	XaaS	Everything-as-a-Service
LS	Least Squares	XAI	Explainable AI
MEC	Multi-access Edge Computing		

II. USE CASES, KEY PERFORMANCE INDICATORS, AND KEY VALUE INDICATORS

One of the key goals and roles of Hexa-X as a flagship project is to identify new UCFs, KPIs, and KVIs, envisioned to pave the way to 6G. The main achievements in this regard are presented in [11] and [26], with more recent updates in [27], where more specific KPIs/KVIs targets are proposed for down selected UCs of each UCF. More specifically, six UCFs have been identified, each of them including several UCs. These UCFs (also appearing at the top left of Fig. 1) are briefly described in the following. *i) Enabling sustainability*: it covers various sustainability aspects, spanning from digital inclusion to environment protection and responsible use of physical/ virtualized network resources, under the umbrella

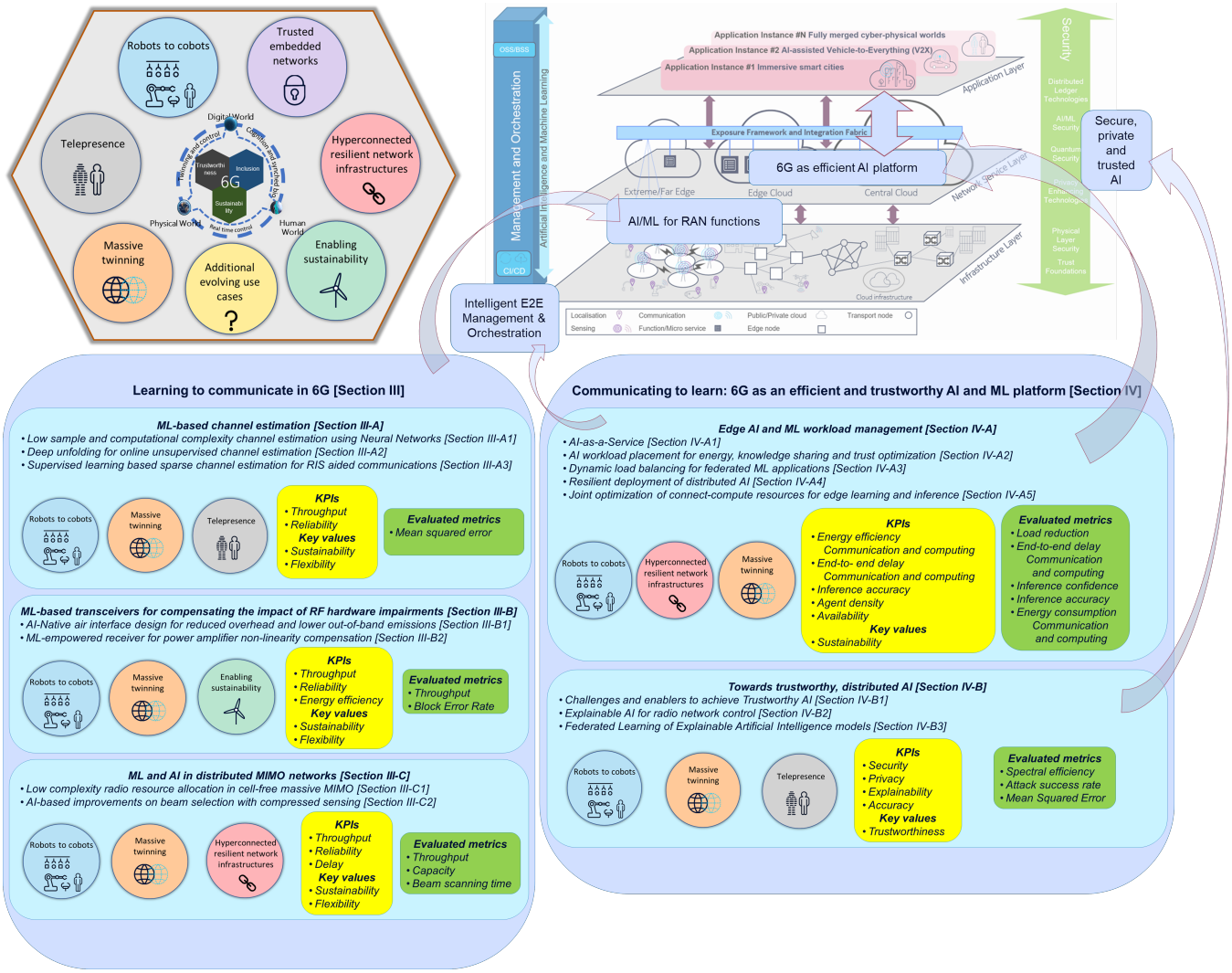


FIGURE 1. Hexa-X use case families, technical enablers and their integration in the architecture, KPIs, KVIs and evaluated metrics

of the United Nations Sustainable Development Goals[3] (UN SDGs). This includes KPIs and KVIs involving, among the others, energy efficiency, trustworthiness, coverage extensions, reliability, and privacy. *ii) Massive twinning:* the scope of this UCF is to enable, in real-time, full digital representations of the physical world. Identified KPIs and KVIs include high link reliability and service availability, low latency, high average and peak data rates, but also, beyond communication, stringent requirements in terms of AI and computing, including agent availability and reliability. *iii) Telepresence*: this UCF aims at enabling the cyber, physical, and digital worlds, to interact with each other in a seamless way from anywhere and at anytime. Among several KPIs, service availability and link reliability, but especially high data rate can be identified, both at communication (transmission) and computation (processing) tiers. *iv) From robots to cobots*: the UCs of this family aim to strengthen the interaction between different types of intelligence (natural, artificial), to enable complex cooperative tasks. Relevant KPIs/KVIs include extremely high communication reliability and low service latency, both at communication and computation tiers. This UCF is of extreme interest for AI and ML related activities. *v) Hyperconnected resilient network infrastructures*: it includes all those services involving (possibly heterogeneous) sub-networks requiring high resilience, e.g., AI-assisted Vehicle-to-Everything (V2X) and AIaaS, with the latter being one of the pillars of the C2L paradigm. Relevant KPIs include very high reliability and low service latency, at both communication and computation tiers. *vi) Trusted embedded networks*: this last UCF refers to all (sub-)network deployments requiring high level of trustworthiness, i.e., with similar KPIs as UCF *v)*.

Although the solutions proposed in this work can be applied to several UCs/UCFs and target KPIs/KVIs, a promi-

[3]United Nations, "Transforming our world: the 2030 Agenda for Sustainable Development," Resolution adopted by the General Assembly, September, 2015, https://upload.wikimedia.org/wikipedia/commons/d/d5/N1529189.pdf

nent role will be played by AI and ML in some of these aspects, with a slight difference between the two paradigms. In particular, as clarified in the sequel, the solutions proposed by the L2C paradigm are more general and applicable to a wide range of UCs, to enhance communication capabilities at different layers of the protocol stack, including throughput, bi error rate (BER)/BLER, channel estimation error (CEE), complexity, spectral efficiency (SE), flexibility, mobility support, energy efficiency, and inference accuracy. On the other hand, the C2L paradigm mostly addresses UCs in which, besides communication, computation plays a key role in the overall performance, aiming to also support the L2C paradigm in network environments characterized by high link volatility. Obviously, the overall vision is to integrate both paradigms into a unified framework in which 6G networks are flexibly and efficiently optimized and automated through learning and adaptation, while providing learning capabilities as a service with challenging performance targets.

Fig. 1 shows an overview of the Hexa-X UCFs (top left side) [26], the architecture (top right side), and the technical enablers described in the technical sections of the paper (Section III and IV). The technical enablers are presented in the bottom part of the figure, following their appearance in this work. As shown in the figure, the work is split into the two paradigms: L2C (bottom left part) and C2L (bottom right part). For each paradigm, different clusters of technical enablers are identified, also corresponding to their specific sections of the paper (indicated in the figure). For each cluster of technical enablers, the addressed UCFs are highlighted by the circles that also appear in the UCFs part of the figure to map them to UCs. Also, the KPIs, KVIs and the (quantified) metrics used to evaluate the performance of the proposed solutions are depicted in the figure through the yellow and green boxes, respectively. As it can be noted, the main key values addressed by our technical solutions are sustainability (e.g., through energy efficiency), flexibility (thanks to ML-based solutions and their generalization capabilities), and trustworthiness (e.g., through explainability and privacy-preserving mechanisms). Finally, the interrelation between the clustered technical enablers and the Hexa-X architecture is highlighted in the figure.

The reader who is interested in a specific topic can easily find it from the figure and refer to the corresponding section. Each section (i.e., the one containing a cluster of technical enablers) is built to be self-contained. However, a compact content overview of the sections as a whole can be found at the beginning of Section III and IV. Also, the reader who is interested in knowing more details about the Hexa-X UCFs, UCs, KPIs, and KVIs is referred to [26], [27]. Finally, more details on the proposed methodologies and technical enablers, as well as specific comparison with state of the art solutions can be also found in previously published contributions, which are referenced throughout the upcoming technical sections. In this paper, we keep a high level description to ease readability while providing a complete overview of the work carried out.

III. LEARNING TO COMMUNICATE IN 6G

In this section, we present a detailed overview of the proposed ML-based technical solutions under the L2C paradigm, with particular focus on ML-based solutions. As summarized in Fig. 1, the technical enablers under the L2C paradigm are clustered into three main groups focusing on channel estimation, RF hardware impairment compensation, and distributed MIMO systems, respectively. They mainly cover the physical layer aspects of channel estimation, air interface design and radio transceivers, PA non-linearity compensation, and beamforming in massive MIMO. While the solutions proposed in this section are largely applicable to any UCs/UCFs mentioned in Section II due to the generalizability of the fundamental physical layer processing tasks addressed, here we downselect several UCs/UCFs which would specifically benefit from the L2C paradigm. The UCFs such as *massive twinning, telepresence*, and *robots to cobots* have stringent requirements in terms of low latency, low BER/BLER, and high data rate, which result in targeted KPIs/KVIs that are hard to accomplish via the existing communications systems and signal processing methods. The proposed AI/ML-based algorithms overcome the algorithmic and modeling deficiencies associated with conventional signal processing algorithms and provide means to achieve these stringent KPI values/KVI levels as discussed in detail in the following sub-sections. The radio transceiver designs for RF hardware compensation proposed in Section III-B also improve the energy efficiency, enabling the UCs related to the *enabling sustainability* UCF.

Furthermore, the ML-based beam selection and resource allocation solutions presented in Section III-C, and the channel estimation solution for Reconfigurable Intelligent Surface (RIS)-aided communication system in Section III-A3 enable novel network architectures that are relevant for UCFs such as *robots to cobots* and *hyperconnected resilient network infrastructures*. They could cater to the *interacting and cooperative mobile robots UC* where a distributed MIMO network could be exploited to manage a cluster of automated (ground and aerial) vehicles over a 6G network. Distributed MIMO architectures powered by AI/ML algorithms would be also beneficial for high-speed vehicular communication in V2X UCs in which extremely high-reliable connections are to be maintained in the mmWave range. Also, such distributed MIMO architectures would be essential to enable the extreme reliability KPIs required by UCs in the *telepresence* and *massive twinning* UCFs.

The next three sub-sections discuss the proposed learning based solutions under the L2C paradigm, presenting technical details and representative results showing how the proposed solutions have achieved the targeted KPIs/KVIs via problem-specific different evaluation metrics, as also shown in Fig. 1.

A. ML-BASED CHANNEL ESTIMATION

Channel estimation is of paramount importance in any communication system. The use of MIMO systems and the

Orthogonal Frequency Division Multiplexing (OFDM) technique in 5G and beyond systems makes it a particularly challenging task as a consequence of the induced complexity. Indeed, traditional techniques often lack behind in this context. For example, least squares (LS) estimation is a low complexity and straightforward one, but it can yield poor performance. An additional example is the minimum mean squared error (MMSE) estimation, which has been shown to give the Neyman-Pearson optimal solution. However, in practice, it requires a lot of computational resources. Furthermore, the sample covariance matrix, which is required for the computation of the MMSE estimator, adds non-negligible overhead to real-time physical layer applications, since a large number of samples are required to reconstruct an accurate sample covariance matrix.

Then, channel estimation is one of the first operations in a radio receiver for which ML-based solutions showed promising results [28]–[30]. In fact, Neural Networks (NNs) have been shown to be particularly suited for the task, typically achieving the best trade-off in terms of estimation accuracy, computational complexity and sample complexity compared to traditional signal processing methods.

1) Low sample and computational complexity channel estimation using Neural Networks

In this section, we focus on NN based solutions for channel estimation that are very computationally efficient. Inspired by the mathematical formulation of the MMSE channel estimator, the authors in [28] propose a shallow NN that obtains close to optimal performance. The resulting NN offers similar performance to the MMSE, with a fraction of computational complexity. However, scaling this NN structure to a larger input pilot size, for instance, to a Massive MIMO system with sounding reference signal (SRS) (a wideband sounding signal) has been proven to be computationally intractable. Inspired by the Kronecker approximation of covariance matrices [31], we propose to consider the dimensions of the channel pilots separately. Our numerical investigations indicate that time, frequency, and antenna space (which can be further divided into horizontal antenna panel and vertical antenna panel) are suitable to split the channel covariance space into subspaces. The choice of the subspaces has been influenced by the following factors. At first, we intend to simplify the pre-processing computational complexity to obtain each subspace. Additionally, for the solution to work properly, we require low statistical correlation among the subspaces. The choice of time, frequency and antenna space meets those two conditions quite well. The resulting NN consists of multiple core NN proposed in [28], [29] repeated and designed for different inputs of spatial, frequency and time domains. Considering two spatial domains (horizontal and vertical domains), we require four NNs. All NNs are similar in architecture, however, trained separately on each of the data subspaces to capture the second order statistics of the data. In the inference time, all four core NNs are cascaded one after another. One of the main

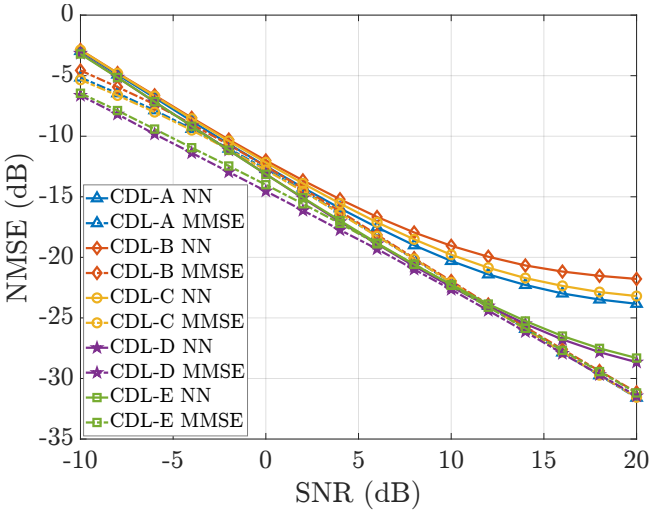


FIGURE 2. Our NN solution tested on different channel data produced from 3GPP CDL-A,B,C,D, and E channel models.

hurdles to actually use a learning-based solution in real world systems is the amount by which the NN solution performance deviates when the channel statistics change over time. Similar issues can arise while training an NN on simulated data for inference on the real data, which is likely to statistically deviate from the synthetic data. In other words, how well it generalizes to channel models which are not in the training data set. We simplified this problem by training the NN on the 3GPP channel models CDL-A and CDL-D while testing it on CDL-A,B,C,D, and E channel models. The numerical results in terms of Normalized Mean Squared Error (NMSE) are compared with the MMSE for each given channel model in Fig. 2. The performance of our proposed method on the models that were not in the training dataset, i.e., CDL-B,C and E seems to worsen especially for higher Signal-to-Noise ratios (SNRs). However, the performance loss in higher SNR seems to persist even for CDL-A, in spite of being in the training set. The model complexity seems to be adequate for the purpose of channel estimation, as the performance loss in higher SNRs is still acceptable.

2) Deep unfolding for online unsupervised channel estimation

As already discussed in the previous section, the complex task of channel estimation is even more challenging in massive MIMO systems, which are at the core of the 5G and beyond systems. In such scenarios, statistical methods that rely on MMSE estimation are at a disadvantage. Fortunately, going beyond the results of Section III-A1, the sparsity of the channels can be exploited along with a physical model to approximate the dominant paths. Indeed, assuming that the channel is a linear combination of a few steering vectors, sparse recovery methods can be exploited to estimate it, provided that a precise knowledge of the physical parameters of the model, such as antennas' positions and gains, is readily

available. This is usually not the case in real-world scenarios where the system has access to nominal approximate values instead. Moreover, hardware impairments, such as those related to frequency generation and acquisition, further complicate the issue. Importantly, [32] shows that a small uncertainty on these parameters leads to high performance loss. Adding flexibility to the estimation model in order to allow it to correct itself and improve its knowledge of the system configuration is thus key. On the other hand, machine learning approaches have gained popularity in the recent years, and have been successfully applied to many domains, with signal processing being not an exception. These approaches are model-agnostic, meaning that they do not require hand-designed models derived from domain knowledge, and rather depend on the quality of the data they are presented with. This however comes at the cost of their computational complexity, as a model's performance is usually correlated with its capacity (i.e., ability to learn complex relationships). In addition, such models usually require huge amounts of data to achieve satisfying results while avoiding overfitting. Recently, model-based deep learning [33] has emerged as a new paradigm aiming at combining the best of both worlds. It consists of exploiting domain knowledge to guide the design of neural networks in order to reduce their complexity and allow them to learn from limited amounts of data.

With this in mind, and going beyond Section III-A1, we propose to use a model-based NN for the task of channel estimation. The NN, called mpNet, is obtained through deep unfolding, which is a technique consisting of effectively unrolling an iterative algorithm so that each layer corresponds to one iteration. The model's parameters (i.e., the NN weights) can be optimized subsequently by learning from data. In particular, we propose to unfold matching pursuit (MP) [34], which is a sparse recovery iterative algorithm.

Two variants of the model are presented for two different scenarios. The first scenario, presented previously in [13], [32] considers a single subcarrier Multi-User MIMO (MU)-MIMO system where a Base Station (BS) equipped with an antenna array communicates with single-antenna users. The NN is initialized with a set of steering vectors constructed based on the current approximate knowledge of the system's parameters modeled by uncertainties on the antennas' positions and gains. It takes as input the noisy channels resulting from LS estimation. The NN training is performed online in an unsupervised fashion, with the objective of minimizing the MSE to the input. This effectively allows a continuous improvement of the a prior imperfect knowledge of the model's parameters. One notable feature is that, since it is observed that the number of iterations of the original MP algorithm depends on the SNR level of the channel, the NN is allowed to have a variable depth, meaning that its number of layers varies for each input according to a stopping criterion.

The second scenario, presented in [35], differs from the first one in that it considers an OFDM Single-Input-Single-Ouput (SISO) system instead. This time, the corresponding model is initialized with a set of imperfectly known frequency response vectors constructed by introducing uncertainties on the subcarriers' frequencies and antenna gains. This is to model hardware impairments such as carrier frequency offset (CFO), sampling clock offset (SCO) and non-flat frequency response of the used antennas. Building on the previous variant, this one exploits two novel ideas, namely constrained dictionaries and hierarchical search. In a nutshell, a constrained dictionary is a dictionary of frequency response vectors where only its parameters (i.e., complex antenna gains and SCO) are allowed to be learned when training the NN, as opposed to learning every entry of the dictionary. On the other hand, hierarchical search is a way of finding the most correlated atom in the dictionary in a hierarchical way instead of the classical exhaustive way, reducing the number of operations to be carried out.

The main metric used to evaluate the performance of this technology enabler is the channel estimation error. Accordingly, both variants are evaluated on synthetic channels in terms of NMSE, as in the previous section. Additionally, they are compared to baselines as shown on Fig. 3 and Fig. 4. We observe that mpNet is able to learn and reduce its error throughout the training and achieves performance similar to what is obtained with a perfect knowledge of the system's parameters.

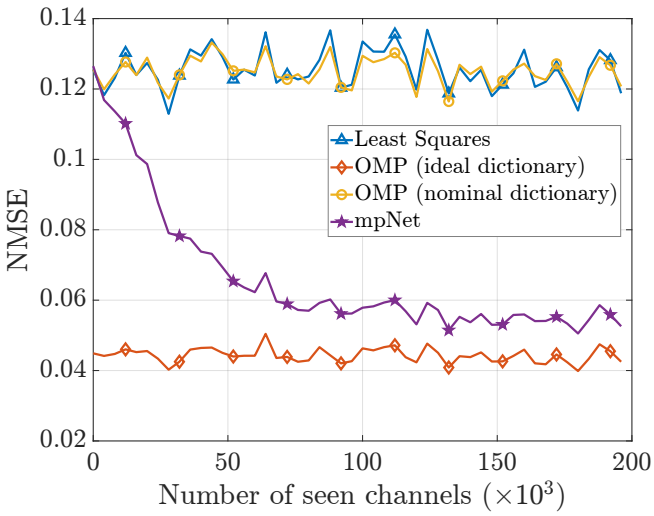


FIGURE 3. mpNet performance in the MU-MIMO scenario.

To summarize, mpNet is a model-based neural network offering the flexibility that classical model-based methods are lacking, while still maintaining a reasonable complexity, unlike classical ML models. Training it without supervision improves its prior knowledge and consequently its performance. It has been applied to MIMO systems and OFDM systems and could be effortlessly extended to systems combining both. Note that a very similar approach has been proposed for integrated sensing and communications [36].

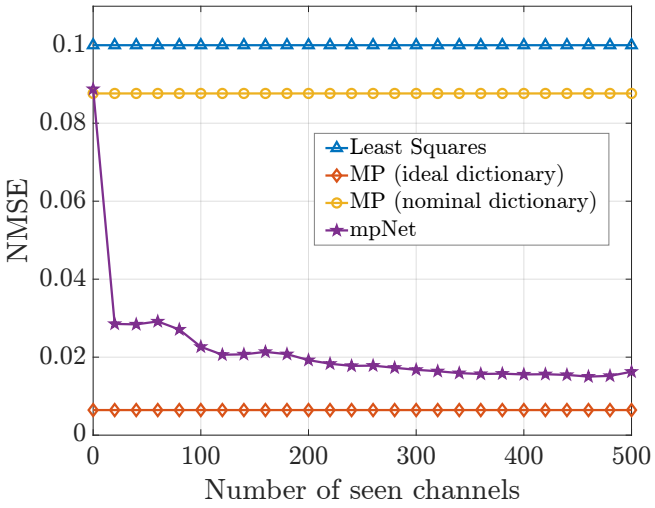


FIGURE 4. mpNet performance in the OFDM scenario.

3) Supervised learning based sparse channel estimation for Reconfigurable Intelligent Surfaces aided communications

The massive deployment of antenna elements does not only pertains to the transmitter and receiver side, but also other nodes in the network than can act as opportunistic reflectors. In this regard, RISs enable the smart control of the wireless propagation environment with software-controlled reflections, but they also introduce new challenges in channel estimation procedures. The scattering of incoming waves can be controlled by configuring the reflection pattern at the RIS, where each element induces a phase shift that can be individually controlled. More and more applications of RISs are proposed in literature [37]–[40] due to the passive nature and lower costs of these devices.

Most of these applications rely on the availability of accurate channel information. However, RISs consist of a large number of passive reflecting elements, which makes channel estimation challenging due to the large dimensionality and the lack of active sensing. In order to reduce the channel estimation overhead, we propose a supervised learning based scheme for the uplink channel estimation of a RIS aided mmWave network [41].

An angular domain sparse channel model is considered by discretizing the angle of arrivals (AoAs). First, the case where AoAs lie perfectly on the discrete grid is considered, and an orthogonal matching pursuit (OMP) [42] based algorithm is proposed to estimate the AoAs. Next, the case where AoAs deviate from grid points is considered, and an NN architecture is proposed to recover the AoAs [41]. It consists of several NNs, where each of them has the signal received at each antenna for all the pilot signals, as input. Further, the complex vector is converted in to a real vector by stacking real and imaginary parts.

Our architecture consists of two parts, where the first network predicts on-grid AoA points using the *sigmoid* activation at the output, which corresponds to the probability of a certain AoA grid point being present. The second network consists of K (i.e., number of AoA grid points) NNs, where each one is in charge of predicting the residual error of the corresponding grid point. It has a *tanh* activation at the output since the residual error can be either negative or positive. Finally, the AoAs are predicted using the output from both networks, and the channel is reconstructed after further estimating the channel gains using the proposed on-grid estimation algorithm.

The performance of the proposed methods are compared against the LS estimator. Fig. 5 shows the performance of channel estimation as a function of the transmit power for the direct channel. As for section III-A2 the CEE (in temrs of NMSE) is used to evaluate performance. We can see that proposed algorithm outperforms the LS estimation, while the NN based approach outperforms both these methods. However, a saturation of performance is seen as the transmit power is increased due to power leakage of the imperfect grid. Fig. 6 shows the CEE performance for the RIS channel, with both proposed methods outperforming the LS estimation with similar performance.

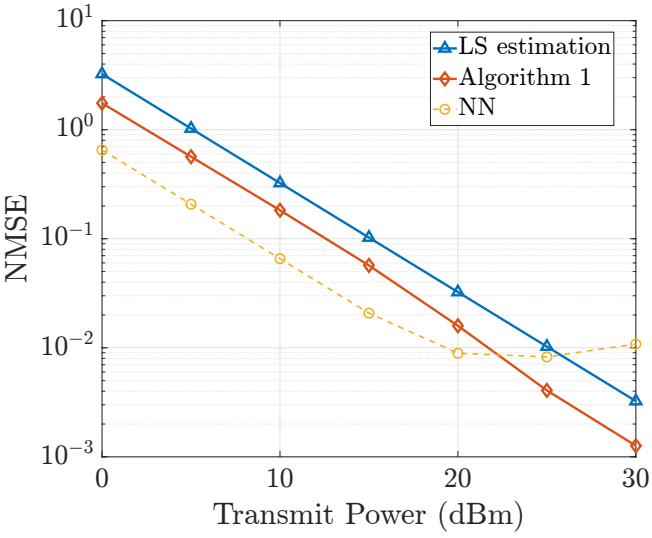


FIGURE 5. Variation of transmit power vs. NMSE for the direct channel.

B. ML-BASED TRANSCEIVERS FOR COMPENSATING THE IMPACT OF RF HARDWARE IMPAIRMENTS

In this section, we focus on the ML-based design of radio transceivers. Designing certain aspects of the air interface to provide native support for AI and ML-based processing is an intriguing prospect for future 6G networks. However, as pointed out but not addressed in Section III-A2, hardware radio frequency (RF) impairments can dramatically degrade the performance of wireless communication. This aspect is not sufficiently addressed in Section III-A. To fill this gap, in this section we discuss how it is envisioned that AI and ML-based techniques are able to effectively compensate the impact of these impairments [43]–[46]. In particular, PA

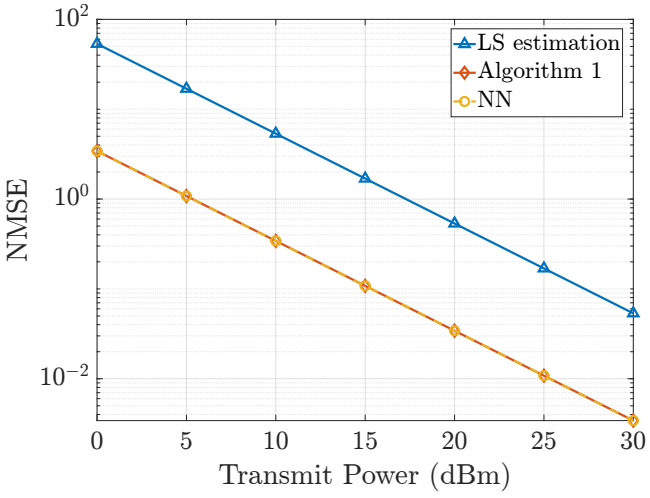


FIGURE 6. Variation of transmit power vs. NMSE for the RIS channel.

non-linearity distorts transmitted signal, causes in-band and out-of-band distortions and degrades throughput of wireless communication systems. Classical methods compensate PA non-linearity at the transmitter-side, e.g., by applying power back-off or performing digital-pre-distortion (DPD). However, applying PA power back-off leads to lower energy efficiency, and lower output power, and hence reduced coverage whereas performing DPD results in high complexity and energy consumption at the transmitter side. AI and ML-based techniques enable compensating the impact of RF hardware impairments by optimizing functionalities at transmitter and or receiver. In the following sections, two techniques are presented for compensating the impact of PA non-linearity: *i)* a method for learning waveforms jointly with the receiver in Section III-B1, and *ii)* a method for learning demapper at the receiver side in Section III-B2, to compensate the impact of PA non-linearities.

As these solutions address the fundamental physical layer processing, they are largely applicable to any selected use case. As for the KPIs/KVIs, the main benefits are in terms of spectral and energy efficiency, thus contributing to sustainability but also flexibility values thanks to the data-driven approach. SE is achieved via reduced BLER and overhead, while sustainability stems from the higher resilience against PA induced nonlinearities, allowing for more energy efficient PA operation. For this reason, as already pointed out, the *robots to cobots*, *massive twinning*, and the *enabling sustainaility* UCFs are identified as the most relevant ones.

1) AI-Native air interface design for reduced overhead and lower out-of-band emissions

An AI-native design might be anything from complete black-box type learning to optimizing the parameters of the air interface, and it has a high potential in improving SE, flexibility, and resilience against hardware impairments, as shown by various earlier works [45], [46]. The black-box type approach entails various challenges when it comes to the practical implementation of such learning-based algorithms, e.g., with regard to the overhead required for training the models during deployment. This means that complete black-box optimization might not be the most favorable approach in terms of performance gain versus system complexity and reliability.

One potential solution for learning-based air interface design without excessive training overhead is to learn the waveform jointly with the receiver algorithm. What is more, only selected properties of the waveform are learned, such as the constellation shape, which means that the system can otherwise rely on, e.g., OFDM waveforms. This also reduces the signaling required for communicating the exploited waveform properties and/or learning them. The primary benefit of such constellation learning is the reduced overhead, as the ML-based receiver can learn to detect the information bits without any pilots when both the constellation shape and the receiver are trained jointly. In addition, the waveform can be made more resilient against nonlinear distortion by learning a Convolutional Neural Network (CNN)-based transformation layer in the transmitter. This can mitigate the impact of the nonlinear distortion produced by the PA, and consequently reduce out-of-band emissions.

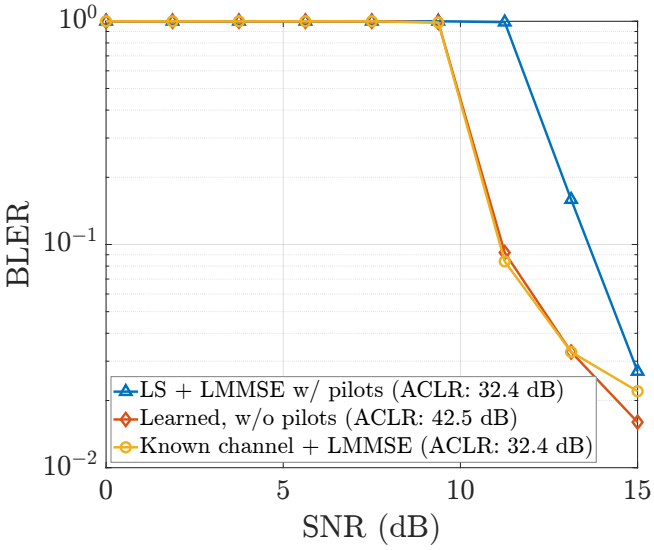


FIGURE 7. BLER of a learned air interface, compared against baseline solutions.

To demonstrate the performance gain of such a learned air interface, Fig. 7 shows the achieved BLER for a sub-THz channel, using a 156 MHz useful signal bandwidth. Three different approaches are considered: an ML-based pilotless air interface with learned constellation shape, a conventional pilot-based air interface with a regular 64-Quadrature Amplitude Modulation (QAM) constellation, and the corresponding performance achievable with perfect genie-aided channel estimates. The learned constellation shape is shown in Fig. 8, where it is evident that a highly asymmetrical shape is learned in order to facilitate pilotless detection at the receiver side.

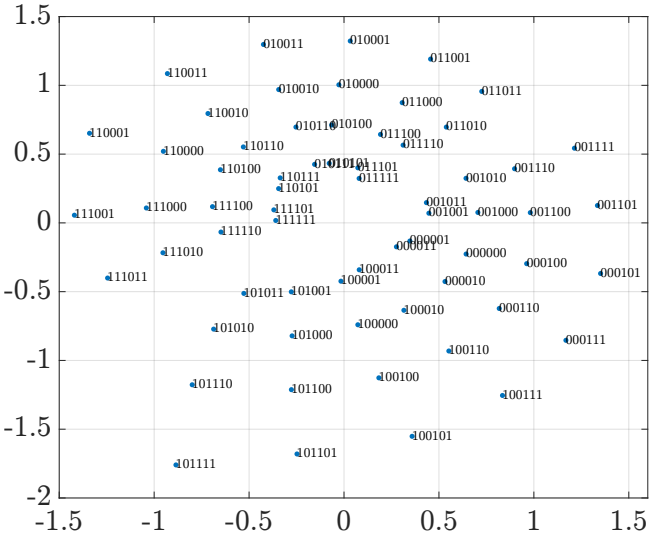


FIGURE 8. The learned constellation shape, showing also the bit mapping.

Firstly, it can be observed that the proposed ML-based approach achieves essentially the same BLER as the genie-aided baseline relying on perfect channel knowledge. With a BLER of 10%, the gain over the practical baseline is over 2 dB. In addition to the BLER gain, the learned solution is also able to achieve higher throughput by not having to reserve any resources for pilot transmission. The achieved adjacent channel leakage ratios (ACLRs) are also shown in the figure, demonstrating that the ML-based approach achieves around 10 dB reduction in out-of-band emissions.

2) ML-empowered receiver for power amplifier non-linearity compensation

Overhead and out-of-band emissions are not the only issues related to hardware impairments. In particular, in high data rate transmission scenarios, the in-band distortions due to PA non-linearity is a limiting factor. In this section, a novel approach is proposed to also compensate the impact of in-band distortions due to PA non-linearity at the receiver side [44]. The developed method can be used towards a wide range of use cases for which improved SE, extended coverage, or enhance energy efficiency are required. This includes high throughput use cases, e.g., Virtual Reality (VR), and Augmented Reality (AR), and the use cases relying on low cost and low energy user equipment (UE).

The proposed method makes use of a neural network-based demapper to compute soft bits to an Low Density Parity Check (LDPC) decoder based on the equalized received signals. Also, it is based on a fully-connected NN operating independently on each resource element. The method can possibly compensate the impact of other hardware impairments, e.g., phase noise as well at the receiver side, as it has been shown for sub-THz transmission in [43]. The developed solution can be deployed either at the base station or at the UE to improve performance in uplink (UL) or downlink (DL) scenarios, respectively. The performance of the proposed method is evaluated and is compared against that of the legacy receiver using link level simulations. Among performance indicators, uncoded BER, BLER, power added efficiency (PAE), and throughput have been quantified for this enabling technology.

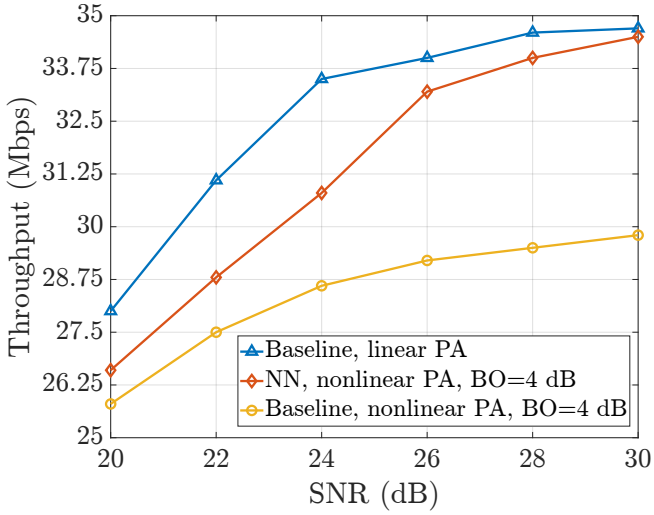


FIGURE 9. Achievable throughput with adaptive modulation order and coding rate at the transmitter for NN-based receiver and legacy receiver for modulation orders up to 64-QAM.

Fig. 9 shows the achievable throughput with the proposed method and with the benchmark method in the presence of link adaptation at the transmitter side with modulation orders up to 64-QAM. The PA back-off is set such that the requirements on out-of-band emissions are full-filled. The legacy receiver in the presence of linear PA provides an upper bound on the achievable throughput. The proposed ML-empowered receiver achieves 17% higher throughput compared with the legacy receiver at high SNR regime. The performance gain would be even higher for higher order modulations where the performance is more constrained with in-band distortions. The proposed method can also enable operation with lower back-off values while achieving similar performance as the legacy receiver, hence, the energy efficiency of the power amplifier can be improved. The simulation results confirm a 70% improvement of PAE for 64-QAM signals. The BER and BLER evaluations confirm that this method can achieve lower BER and BLER and hence improve the reliability of communication links. It has been shown that in certain cases, e.g., higher order modulations or higher code rates, this method can provide a reliable link. For instance, it can reach certain levels of BLER (10%) where the legacy method fails to provide a reliable link.

The proposed method can be deployed in different scenarios such as the uplink communication of a cellular network, which is usually coverage limited. In this case, performing linearization techniques at the UE is challenging due to the limited processing capability and energy budget of the UE, and it is desired to enhance the UE energy efficiency to

increase its battery lifetime. Alternatively, this method can be used in downlink scenario to increase throughput and/or extend the coverage area of the base station, and to relax the requirements on DPD for in-band distortions, especially for high throughput transmissions, and to improve the energy efficiency of the base stations. The improved energy efficiency of the base stations can lead to smaller size and weight due to the reduced requirements on cooling equipment.

C. ML AND AI IN DISTRIBUTED MIMO NETWORKS

Massive MIMO is a key technology component for future wireless networks, since it provides high beamforming gain and leads to increased spectral and energy efficiency. Recently, distributed massive MIMO (D-MIMO, or, cell-free MIMO) systems are extensively investigated as a potential MIMO architecture, where a large number of distributed accesss points (APs) are connected to a central processing unit (CPU) via fronthaul links to serve a much smaller number of users distributed over a wide area, and using the same time-frequency resources, without classical cells or cell-boundaries. These architectures provide more uniform service performance for the users in terms of SE, but also connection robustness due to the additional spatial diversity. For this reason, they are extremely relevant for the *interacting and cooperative mobile robots* UC (as part of the *robots to cobots* UCF), where new cell-free massive MIMO architectures could be used to manage a cluster of drones over a 6G network along with the novel ML-based resource management algorithms described in the following. The complexity gain and flexibility are the main KPIs evaluated for these enabling technologies, thus contributing to sustainability and flexibility values.

However, the increased number of antennas in the system heavily increases the complexity and overhead of the optimization problems to solve, highlighting the need to have efficient and scalable solutions for emerging tasks. Resource allocation performed in a system with coordinated operation of many APs makes traditional optimization-based solutions infeasible. The problem is addressed in Section III-C1 by proposing a data-driven scheme to perform joint power and fronthaul capacity allocations with decreased complexity. Another overhead appears in analog beamforming systems due to beam selection, which becomes severe in D-MIMO networks with a large number of beamforming APs. Compressed sensing is a promising technique to reduce the beam selection, which is further optimized by a learned dictionary and neural sparse decoders, as described in Section III-C2.

1) Low complexity radio resource allocation in cell-free massive MIMO

In cell-free massive MIMO networks, proper radio resource allocation such as power control and efficient utilization of the limited fronthaul links is essential in achieving improved performance. However, apart from the computational complexity issue mentioned above, conventional optimization or heuristic-based algorithms face several challenges such as sub-optimal solutions in complex and non-convex problems, lack of flexibility and parameter sensitivity, and inaccuracy of the model-based resource allocation methods [47]. In recent literature, the capability of ML-based algorithms has been exploited to overcome those challenges associated with conventional approaches. The works in [48]–[50] proposed ML-based power control algorithms for massive MIMO networks via supervised learning where a deep neural network (DNN) is trained to learn the mapping between the inputs (user locations or channel statistics) and the optimal power allocations obtained by an optimization algorithm. On the other hand, our previous work in [51] proposes to learn the power allocation in an unsupervised manner by training the DNN over the optimization objective.

In this section, the unsupervised learning approach in [51] is extended to learn joint resource allocation tasks in a cell-free massive MIMO network. Specifically, joint optimization of user power allocations and fronthaul capacity allocations, between Channel State Information (CSI) and data, to maximize the network sum throughput in the uplink of a limited-fronthaul cell-free massive MIMO network is considered. The system model and problem formulation are similar to [52] and an ML-based algorithm is proposed to solve the sum rate optimization problem instead of the geometric programming-based solution or weighted minimum mean squared error (WMMSE) approach. In order to solve the joint optimization task, a DNN which we denote as *PowerNet_Ext* is directly trained using a custom loss function to optimize the sum throughput objective. The large-scale channel coefficients between the users and the access points are used as the DNN input and the DNN is trained to output the user power allocations and fronthaul capacity allocations between the APs and the CPU to maximize the system sum rate. The performance of the proposed algorithm is evaluated for a cell-free MIMO system with 50 APs and 10 users distributed in a simulation area of 1×1 km^2. The estimate-compress-forward (ECF) strategy is considered for the CSI and data transmission between the APs and the CPU. With the ECF strategy, once each AP receives the pilot and data signals, it performs the MMSE channel estimation and then separately quantizes the estimated channel coefficients and data signals to forward them via the fronthaul link. Other simulation parameters, data set preparation, and the model training procedure are similar to [51].

Fig. 10 compares the sum SE performance of *PowerNet_Ext* and the geometric programming-based algorithm proposed in [52] for the ECF strategy, for the cases of perfect transceivers and with transceiver hardware impairments. The *PowerNet_Ext* achieves close performance to the baseline performance in both perfect and imperfect hardware scenarios. An online training stage is also introduced exploiting the unsupervised learning capability of the DNN. In this case, during the inference stage, the originally trained model is retrained for several iterations for each channel realization. Performing online training allows further customization and fine-tuning of model parameters based on large-scale channel

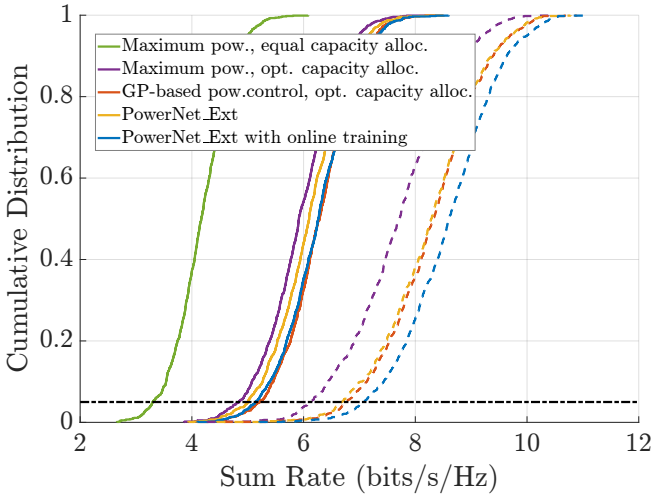


FIGURE 10. Sum rate comparison between *PowerNet_Ext* and the baseline [52] with joint power control and fronthaul capacity allocation for 50 APs and 10 users for ECF strategy. Total fronthaul capacity C_m = 1 bits/s/Hz. Solid lines: with transceiver hardware impairments ($\kappa_t = \kappa_r = 0.9$) and dashed lines: perfect transceivers ($\kappa_t = \kappa_r = 1$).

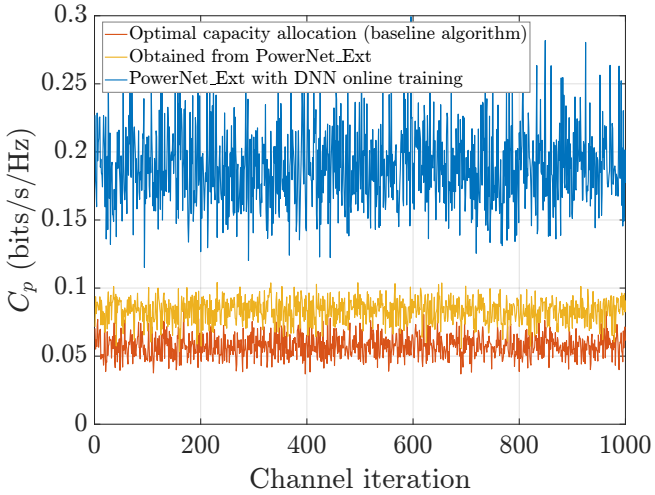


FIGURE 11. Variation of optimal CSI fronthaul capacity (C_p) allocation over channel realizations, obtained from *PowerNet_Ext* and as proposed by [52].

inputs in each channel realization to further optimize the sum rate performance, as it can be seen from Fig. 10. Furthermore, Fig. 11 depicts the obtained optimal CSI fronthaul capacity allocations between each AP and the CPU as obtained from the three methods when the total fronthaul capacity of each AP is C_m = 1 bits/s/Hz. According to the one-dimensional search algorithm proposed by [52], all the APs are allocated the same fronthaul capacity C_p for CSI transmission, however, *PowerNet_Ext* is capable of learning different C_p values for each AP depending on its channel conditions which help to improve the sum rate performance as seen from Fig. 10.

The geometric programming algorithm in [52] for uplink power control has an $\mathcal{O}(K^{7/2})$ algorithmic complexity that scales with the number of users K [53]. In contrast, the *PowerNet_Ext* only does a one-shot calculation performing a series of matrix multiplications and additions and function mappings in each layer to produce the outputs and hence has a fixed algorithmic complexity. Thus, the above numerical simulations show the potential of the proposed ML-based approach in learning resource allocation vectors resulting in similar sum throughput performance compared to an optimization-based baseline, while having lower computational complexity.

2) AI-based improvements on beam selection with compressed sensing

Keeping the focus on massive and D-MIMO settings, beamforming is a very efficient technique to provide reliable coverage at higher frequencies, but it also poses a significant challenge to the beam selection process in the form of increased scanning overhead. High delays in beam selection have especially large impact on connection reliability in applications with high user mobility or in deployments with many blockages. Independently from beamforming, Compressed Sensing (CS) theory is a relatively newly explored scheme, which has found its application in different areas recently, e.g., imaging applications, radar signal processing, but also in wireless systems. CS is a signal processing technique, which states that if a signal is sparse in some domain then it can be reconstructed from fewer number of measurement samples than what would be required by sampling theory. Being sparse means that the signal contains only a few non-zero elements if it is expressed on a certain linear basis.

In case of mmWave systems, the radio propagation on multipath channels is known to be sparse in the angular domain. We consider a system with n transmit antennas forming n equally spaced beams with a DFT-based codebook and one receive antenna at the UE. The reference signal for beam selection is transmitted over m time slots with a certain power distribution over the beams according to the dictionary (or sensing matrix) $\mathbf{D}^{(m\times n)}$. As the reference signals are linearly combined at the receiver, we obtain a standard sparse decoding problem of $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $\mathbf{y}^{(m\times 1)}$ is the received signal and $\mathbf{x}^{(n\times 1)}$ is a sparse vector of the channel in the angular domain. This problem can be solved even if $m \ll n$.

This dictionary training can be performed with the help of an autoencoder architecture, where the input is a representative set of beam channel vectors $\mathbf{x}$ sampled from potential UE locations. The encoder is a dense layer (matrix multiplication) with the dictionary elements $\mathbf{D}$ set as weights, which results in the measurements $\mathbf{y}$. The decoder block must implement a sparse decoder algorithm using the encoder dictionary D and it must also be differentiable so that gradient descent can be applied to optimize $\mathbf{D}$. Several neural sparse decoder architectures are investigated based on the Learned Iterative Shrinkage and Thresholding Algorithm family of LISTA ([54], [55]), resulting in a recurrent NN decoder block.

The analysis is performed based on the open DeepMIMO dataset [56], with 4 access points and 32 horizontal beams on each of them, which means 128 beams altogether with 128 measurements if sequential scanning is used. Fig. 12

shows results for 3 solutions. The first solution uses random dictionary and iterative sparse decoding using a conventional least absolute shrinkage and selection operator (LASSO) algorithm. The other two scenarios show the results with a dictionary trained for the local environment and sparse decoding performed with either the LASSO optimization or the fully NN-based sparse decoder LISTA. The key KPI to compare the different solutions is best beam matching ratio within 1 dB (the decoded beam is accepted as best beam if its gain with path loss is within 1 dB compared to the best). Note that the neural sparse decoder has its own set of trainable parameters which is dependent of $\mathbf{D}$ but further optimizes the decoding process with better accuracy and less iteration layers required. Although the learning was performed jointly, $\mathbf{D}$ can be used with any other traditional sparse decoding algorithm. However, the efficiency of the proposed solution can be further increased by applying the neural sparse decoder. In the investigated scenario, the required number of measurements due to standard CS with random dictionary at 90% targeted KPI is reduced from 128 to 32, which is improved further significantly by using optimized $\mathbf{D}$ and also an optimized decoder.

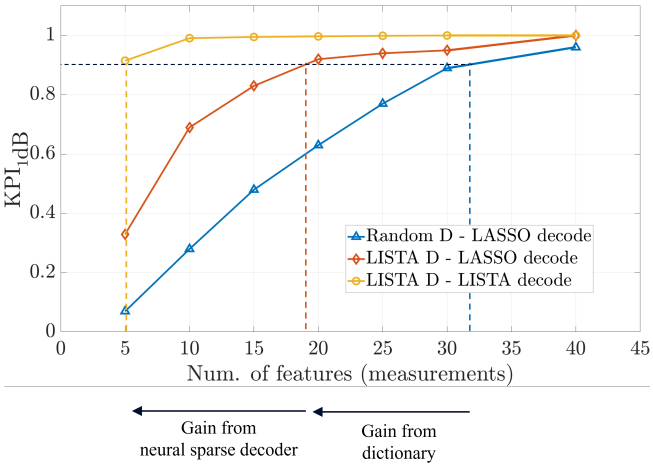


FIGURE 12. Observable scanning time gains compared to the baseline CS-based sparse detection

Decreasing the beam scanning time is especially beneficial in beam tracking situations when fast beam changes are required to prevent losing connection for latency critical communication. Significantly less scanning overhead can easily be translated into more frequent beam updates, leading to both lower ratio of connection drops as well as faster recovery times.

IV. COMMUNICATING TO LEARN: 6G AS AN EFFICIENT AND TRUSTWORTHY AI AND ML PLATFORM

Beyond the role of AI and ML as enabler for flexible network optimization, as described in Section III, there is an increasing adoption of intelligent components among higher-layer in-network functions and external applications. For instance, recalling the *interacting and cooperative mobile robots* UC, it can be noted that it requires real-time intelligent decisions based on distributed and resource efficient data and model sharing. Similarly, in applications of the UCF *hyper-connected resilient network infrastructures*, a huge amount of data should be distributed across thousands or millions of devices. Instead of sharing high volumes of raw data, which may not be feasible due to communication, capacity, privacy, complexity, and other reasons, optimized neural encoded/embedded data representations can be exchanged to guide data-centric decisions. Finally, the specific requirements of distributed AI deployments should be taken into account for the *digital twins for manufacturing* use case (as part of the *massive twinning* UCF) during workload placement, where data availability and trust levels are also taken into account.

These requirements call for a joint communication and computation co-design, leading to network services and Application Programming Interfaces (APIs) with seamless exploitation of network knowledge for both in-network and external applications. The challenges of the wireless environment, energy efficiency, device capabilities and data handling constraints require 6G networks to provide efficient platform support for distributed AI learning and inference functions.

This section provides a detailed overview of two large groups of technical enablers of the C2L paradigm. Edge AI and ML workload management in Section IV-A targets the KPIs of energy efficiency and end-to-end (E2E) application delay, by accounting for both communication and computing components in the processing chain. These performance metrics contribute to the KVIs leading to sustainability values. AI agent availability and inferencing accuracy are also shown to be supported in high-mobility environments involving safety-critical communications. In AI workload placement multiple KPIs are considered, including AI agent availability, network energy efficiency (by targeting reduced energy consumption), as well as trustworthiness (by prioritizing trustworthy physical nodes). The technical enablers of trustworthy, distributed AI in Section IV-B are more focused on the KPIs belonging to the trustworthiness value, whose KVIs can be summarized into security, privacy and explainability, with also a focus on maintaining the model accuracy at a prescribed level.

A. EDGE AI AND ML WORKLOAD MANAGEMENT

In-network and external AI and ML related workloads will be more and more pervasive in 6G systems. Natural questions arise on several aspects. First, AI and ML models should be provided to client devices as a service and upon request, which needs a new paradigm known as Compute/AI-as-a-Service. Then, once the network is able to offer these capabilities, one of the first issues is where to place such workloads, with metrics including energy, end-to-end delay (including communication and computation), and learning/inference accuracy. This covers single client workloads, with possible distribution across distributed heterogeneous nodes, but also FL settings, in which load balancing also plays a key role, depending on devices' availability and

data distribution. Once the various workloads coexisting in the network are placed, resource allocation of wireless and computing resources is fundamental to strike the typical edge learning trade-off between energy, delay, and accuracy. This section covers all these aspects, from the foundations of the AIaaS concept, to distributed inference and the joint allocation of radio and computing resources for edge learning and inference workloads. As already introduced, several UCs of in different UCFs are covered by the proposed solutions, especially those entailing computing resources in the network as means for performing complex (cooperative) tasks.

1) Artificial Intelligence-as-a-Service

The integration of AI for internal network operation (e.g., resource allocation optimization, etc.) as well as to external entities is envisioned to take place through an AI-as-a-Service approach. This is a relatively new field of research. [57] provides an overview of existing works and explains that concepts for moving AI processing into the cloud or network Edge are still under intense investigation. A specific example of a related application is examined by [58], which introduces a configurable model deployment architecture for Edge AIaaS, which enables a management entity to optimize the energy and delay performance by jointly customizing the task data quality, model complexity and resource allocations with given Quality of Results (QoR) constraints. The approach is applied to an optimization model that minimizes the hybrid energy-delay cost by jointly optimizing task configurations and computation resource allocations. Furthermore, the application of related concepts is studied in specific fields such as autonomous driving [59], facial recognition [60] and others. In the context of Hexa-X, the overall concept is being broadened by introducing open interfaces to AI Services in the network being made available to the network itself as well as external consumers of AI Information Services (AIS). The objective is to make the knowledge of the network available to the benefit of all, while protecting the owner of the underlying data as well as the privacy of the AIS consumers. This objective is achieved by users requesting the derivation of a learning model optimized for a specific and well defined purpose. The network is then able to fully exploit its knowledge and to transfer the final trained learning model to the users. Hexa-X has published its solutions, for example outlining an approach to an AI Architecture in [61] and showcasing how Hexa-X can further support the implementation of European Regulation in the the European Telecommunications Standards Institute (ETSI) white paper [62].

To be more specific, the AIaaS approach will enable an AIS consumer (via a User Interface) or a client application to request a parameterization of the device's locally available learning model from the network as illustrated by Fig.13.

The network will derive the requested learning model given a number of performance requirements set by the user, the client/server application or a UE profile. Whether direct (involving a subscription to the selected AI agent(s)) or indirect AIS consumer (e.g., UE) and AI agent communication is better applicable depends on: *i)* the considered scenario - whether it involves a single one or periodic/ frequent inferencing-based decisions, and *ii)* whether the UE and the selected AI agent can communicate via a common application layer protocol. For example, in case a single prediction is needed, the indirect communication case may be better as there is no need to subscribe to ML model updates. However, in the case of e.g., Quality of Service (QoS) prediction for a given vehicle trajectory in the context of autonomous driving, subscription to AI agent model updates may be needed as multiple predictions may need to be performed (e.g., for different parts of the route or even more fine-grained predictions referring to the same waypoint).

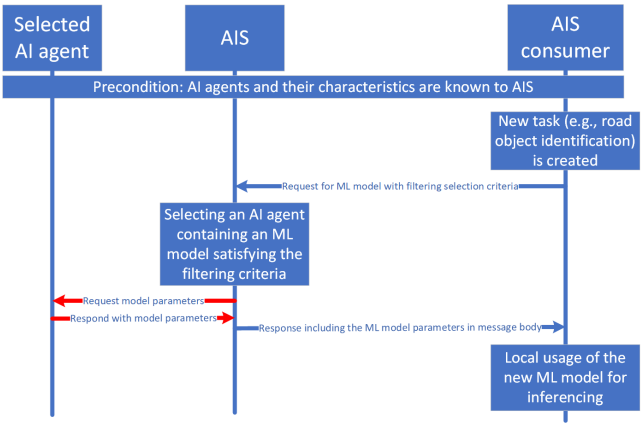


FIGURE 13. Signaling flow for requesting and delivering of a new training model satisfying AI agent selection criteria posed by the AIS consumer (e.g., UE). Subscription-based direct UE/ AI agent communication case - e.g., for frequent/ periodic inferencing-based decisions.

Unsubscription from an AI agent or subscription updates may be needed in case e.g., the UE moves away from the network entity (e.g., Multi-access Edge Computing - MEC host) hosting the AI agent. In this case, the AIS needs to be contacted again with updated filtering criteria, in order to target AI agents hosting models relevant to the problem/ task, and that can provide their updated ML models with low latency.

Consequently, full knowledge of the network can be seamlessly exploited for resolving the specific problem stated by an AIS consumer without exposing the network data sets directly. The solution is applicable to any type of commercial or professional applications, including safety and dependability-critical environments (automotive, industrial automation and others).

The proposed AIaaS approach will be implemented through an AIS and its corresponding AI Application Programing Interface (AI API), exploiting an open network interface. The proposed service and API enable the following:

- A UE (AIS consumer) to communicate to the AIS information relating to a user/ client application-specific task (e.g., intention to drive a vehicle from location A to location B, starting at time t) calling for an inferencing-

based recommendation (e.g., QoS prediction-based recommendation on switching on/ off autonomous driving features) and performance requirements relating to e.g., inferencing accuracy, energy efficiency, end-to-end delay, security and others. All these criteria are filtering criteria for AI agent selection.
- The UE, based on AIS response on available AI agent(s) fulfilling the communicated criteria, to *i)* in case of a commonly supported application layer protocol, subscribe to, unsubscribe from or update the subscription to one or multiple available AI agents (e.g., FL aggregators), or *ii)* in case infrequent/one-time output is needed to obtain the ML model configuration indirectly from the AIS.
- Considering each selected AI agent, the UE to share its local model updates to the AI agent(s) it is subscribed to and obtain learning system parameter updates (e.g., aggregated FL model update, transfer of an already trained and tested model) by the subscribed AI agent(s).

The AIaaS approach offloads the complex and computational resource intense process from the client device to the network. Once the desired learning model is created by the network and forwarded to the client device, limited in-device capabilities are required, for example a neural network accelerator component, in order to apply the learning model and take full advantage of the full knowledge of the network.

2) AI workload placement for energy, knowledge sharing and trust optimization

AIaaS is a powerful and novel concept that allows users to access network intelligence and knowledge on demand. However, it also introduces several challenges, among which the workload placement represents the first step towards a trustworthy and sustainable decision-making process. Hence, managing the AI operations in decentralized scenarios, i.e., where multiple nodes/devices may participate in the execution of diverse AI workloads, is crucial.

There are many recent studies in literature related to optimal Virtual Network Function (VNF) and service placement in beyond 5G (B5G)/6G networks, and a limited number related to AI workloads placement. The work in [63] proposes a reinforcement learning framework with an efficient representation and modeling of the state space, action space and the penalty function in the design of the underlying Markov Decision Process. The aim is to minimize the network delay and the number of edge servers and provide a MEC design with minimum cost. The authors of [64] propose an AI-driven online policy called SplitPlace that places neural network split fragments on mobile edge devices using decision-aware reinforcement learning. The aim is to fine-tune the placement of computing tasks in volatile environments. Moreover, [65] tackles the problem of energy-efficient virtual security functions (VSF) placement to minimize energy consumption while meeting flow-level security requirements and resource constraints. The problem is formulated and solved with an integer linear programming (ILP) model which minimizes server energy consumption, and it also proposes a heuristic algorithm for large scale network instances. Finally, optimized sequential Service Function Chains (SFCs) placement is performed in [66] with the development of a delay and location aware Genetic Algorithm-based approach for minimizing the end-to-end delay of ultra-low delay industrial network operations and also exploiting location aware information.

Here, a novel AI workload management system is proposed, applicable in B5G/6G architectures for close to optimal mapping of AI workloads to the various network's physical nodes (e.g., user devices, edge/cloud servers). Physical nodes that undertake the execution of AI workloads can face trust level problems, traffic load-related issues, or energy consumption problems; to this end, an optimization algorithm is designed and developed, which targets a threefold strategy, i.e., to minimize the power consumption of the overall network towards sustainability, to minimize the processing and transmission delay, and to maximize the overall trust level of the system, by prioritizing nodes/AI agents with high trustworthiness indexes. Hence, the objective function (OF) that is being minimized consists of three terms as described before, to which respective weights are applied (w_1, w_2, w_3) depending on the use case requirements: $\text{OF} = w_1 \cdot P + w_2 \cdot D - w_3 \cdot T$. The power consumption term P is the sum of the respective power consumption of each physical node/server and is mainly affected by CPU/GPU/NPU utilization rate compared to disk storage, memory and bandwidth [67]. Subsequently, the delay term D consists of the sum of the respective transmission and processing delay of each AI workload and the trust term T consists of the sum of the trust level index of the physical nodes/servers used for the placement.

The described problem is solved with the development of a meta-heuristic algorithm building upon the genetic algorithm paradigm [68]. The proposed algorithm includes several optimization steps, including: *i)* a dynamic stopping criterion for faster convergence and termination compared to the one used in the classic version of the genetic algorithm, *ii)* an efficient initialization of "chromosomes" (particular solutions) so that a feasible solution is ensured when the system is close to fully loaded, *iii)* a penalty function for computational requirements constraint-handling, and *iv)* an efficient mutation form (suitable for the non-binary "chromosomes" utilized) for exploring the whole search space.

The performance of the developed algorithm was compared with the output of a Mixed Integer Programming (MIP) solver [69]. The proposed genetic algorithm obtains close to optimal scores within significantly less time than the MIP solver as the number of AI workloads increases. Specifically, when the number of physical nodes is 43 and the number of AI workloads exceeds 80, the MIP solver performance is intractable, compared to the proposed solution, which demonstrates an execution time of approximately 38 sec. Moreover, the percentage of reduction of processing and transmission delay was measured as the number of AI

workloads increases, for three different weight levels w_2 (low, medium, high). In order to assess the effect of the delay-specific OF term, we compare the optimization results (Fig. 14) for $w_2 = 0$ with three $w_2 \neq 0$. The larger the weight of the delay term w_2 is, the higher reduction of delay we observe. Additionally, it is observed that for higher number of AI workloads, i.e., more than 150, the reduction of delay decreases. This is due to the fact that for such cases, the number of feasible placement solutions decreases and potential gains in the reduction of delay are limited.

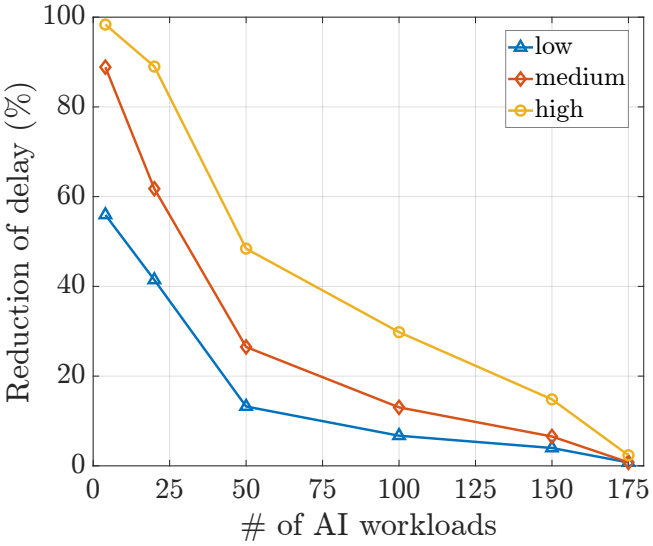


FIGURE 14. Percentage reduction of processing and transmission delay with increasing AI workloads for different weight levels (w_2). Baseline approach refers to the proposed genetic algorithm having $w_2 = 0$.

Future work involves the quantification of the trust level index as well as the elaboration on the architectural implications and respective requirements.

3) Dynamic load balancing for federated ML applications

Workload placement and sharing becomes even more challenging in several envisioned 6G use cases in which a large number of heterogeneous sensors are connected to FL nodes at the network edge. For example, massive twinning relies on a fully synchronized and accurate digital representation of the physical and human worlds based on a wide variety of sensor information. To enable more efficient interaction of production for digital twins in manufacturing, we have to encompass a larger extent of the respective processes, and also to achieve the transfer of massive volumes of data from a wider range of sensors and actuators within the factory, including the cooperation among multiple digital twins in a flexible production process. In an immersive smart city, effective management of all factors of persons, vehicles, infrastructure, weather, pollution, etc., on various time scales require a large number of heterogeneous sensors to be connected.

Given a large number of edge connections, load balancing is necessary to remedy potential hot spots and data diversity to ensure quality balance for the federated learners. By dynamic load rebalancing, we reconnect sensors to nodes in the radio network if load is uneven or some nodes receive insufficient variety of data for serving local models.

In the use case scenario of Fig. 15, a variety of sensors are connected to AI compute nodes through radio BSs. For accurate and low latency operation, each AI node needs access to sensors of most types, and the connection load needs to be balanced. Based on the Timing Advance (TA) information, the connection of sensors to AI nodes can be reconfigured; however, in addition to handover costs, the state of a sensor may also need to be migrated to the new AI node.

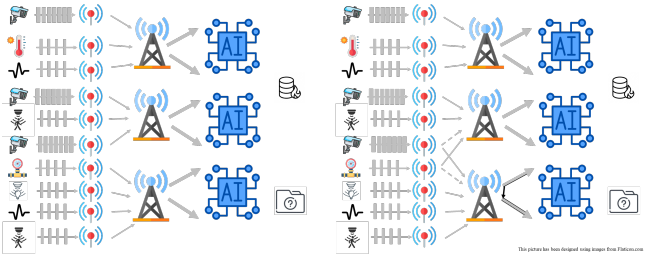


FIGURE 15. Left: a FL hot spot with too much data (top) and an insufficient data with one type (camera) missing (bottom). Right: a reconnection decision causes state migration between the bottom AI agents.

Our goal is to provide load balancing to remedy potential hot spots and data type diversity to ensure quality balance for the federated learners. Load and diversity balance is necessary to make sure each node can equally contribute to the FL task dynamic load rebalancing by reconnecting sensors to nodes in the radio network if *i)* load is uneven, or *ii)* some nodes receive insufficient variety of data for serving local models, for example when a certain crucial type of sensor is not connected to a FL node.

Our proposed dynamic reconnection solution is based on the Key Isolator Partitioner (KIP) originally developed for distributed data processing systems [70]. KIP is a heuristic combination of explicit placement and weighted hash partitioning to improve balance in cases of heavy data skew. KIP involves a distributed top-k histogram computation, where locations with heaviest load are ordered by decreasing frequency in a histogram object. The ideal maximal load of the partitions is calculated using a soft threshold to guarantee a good balance. In KIP, first the highest load is arranged greedily by considering radio accessibility. KIP attempts to keep UEs in their current connection to minimize migration costs, and non-heavy keys are handled by the weighted hash partitioner. The average load of a node is computed, and the UEs are rerouted as necessary by greedy bin packing. KIP prepares for potential reconnection by making minimal modifications to the existing network state.

4) Resilient deployment of distributed AI

When AI workloads are integrated into wireless networks, not only computing management is required, but rather a

holistic view of communication and computing is envisioned to lead to significant gains in terms of resource efficiency and resilience. This pertains to several UCFs with application systems highly distributed on many different devices and network components. These AI-enabled components will jointly realize a heterogeneous AI and data sharing landscape where stored data, real-time sensor input, processing and control capabilities are distributed, AI components may be integrated on multiple levels with full or partial AI processing forming loosely or more tightly coupled systems, and the nature of shared data is similarly heterogeneous covering raw sensor inputs, model states, latent spaces with implicit or explicit semantics.

Use cases realizing real-time critical functionalities in such distributed environment would pose stringent requirements on the communication network in terms of packet latency, loss rates, bandwidth stability, device density, along with a high signaling overhead to manage it over wireless. However, low latency, high reliability and availability of AI-enabled applications can also be realized by combining AI-level resilience techniques with supporting communication functions in 6G. In this concept we exploit two commonly occurring properties: *i)* a certain level of redundancy in input data among different sources and *ii)* achieving incrementally increasing accuracy/reliability by extending on input collection time. These properties allow AI applications to cut the long tails stemming from highly variable radio environments, while at the same time relaxing on the individual communication link parameters.

We consider a scenario of distributed sensing and communications where the application performance benefits from tight integration of sensors and communication. A large number of devices are deployed in a wireless environment and are equipped with sensors providing input for intelligent fusion, realized by a joint inference engine running at the edge. One example application is the cooperative perception, one of the advance use cases in the 5G Automotive Association (5GAA). It involves sharing sensor information about the current driving environment among the vehicles and other roadside stations. Using sensor data from nearby objects allows the participating vehicles to increase accuracy of the estimated parameters and form a more complete state of environment, including, e.g., blocked objects. However, this shared sensor data can be highly redundant, with a noisy input and variable link quality. Sharing all inputs with the inference function will likely require unnecessarily high bandwidth usage, thus limiting the main target KPIs: inferencing accuracy, inferencing latency and device density. The device scalability potential with traditional sensor fusion is also limited due to potential communication and computing bottlenecks in the network.

To solve the above problems, an AI application and communication system architecture is proposed. The inferencing application is based on an AI framework, which can perform early inference from partial data for low latency applications, simultaneously fulfilling high accuracy requirements with increased delay. The contribution of the individual input sources to the joint inference task is also varying due to the observation noise (input image quality, visual blockage, etc.) and communication link quality. The solution can also prioritize among the inputs based on added value, which can decrease the required data volume over the shared wireless channel, allowing higher device density to be served.

The above requirements can be supported by multiple AI architectures. This study investigates the case with spatio-temporally trained Spiking Neural Network (SNN) [71], which has the properties suitable for these goals. Although this family of neural networks was developed for SNN, it can be implemented as a stateful Artificial Neural Network (ANN) using discretized time slots. The resulting network will be sparse in communication, with only spike-type data transfer among the neurons (1 bit in the discretized implementation) with very low activity level. The inference decision in an object recognition task is performed by accumulating the spikes in the final evaluation layer, which can effectively be translated into an ordinary logit layer. The advantage of this architecture is that both the noise from input sensors and the information loss on the wireless link can be offset by increasing the inference time. This mechanism provides less accurate but low delay results as well as incrementally increasing the accuracy to the required level with increased latency. In this way, the application has the ability to control the inferencing process in accordance with the actual application level targets.

It is also assumed that the inference control function in the application can prioritize between different input streams. This partial evaluation is performed after each time step. An input stream utility assessment is made, which may stop a device sending a stream or adjust communication bandwidth among the live streams according to data utility to increase inference accuracy. Significant gains can be observed in the wireless communication load, both in terms of traffic level and number of active connections (Fig. 16). By gradually eliminating redundant input streams, the average traffic load decreases to less than 20% compared to the baseline case and the average number of required connections is also in this range.

The proposed joint compute and communication system has several advantages over the traditional approaches. With the option of accuracy-latency trade-off, it is possible to do both early phase ultra-low latency inference, as well as higher accuracy at the cost of higher delay. In use cases like above, where multiple inputs provide overlapping information for the inference task, over-the-air communication can be significantly reduced with the help of application domain functions (the assessment of input stream quality and utility) interfacing with network layers on a fast and high data granularity level, which may require per packet control on millisecond timescale. The benefits, however, are in the order of $1/5^{th}$ of traffic and connection load, which can also be translated to higher device density for distributed AI.

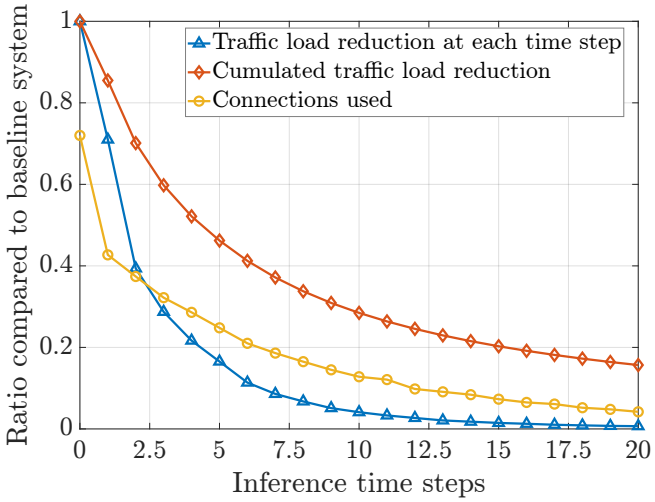


FIGURE 16. Incremental inference for an ANN-SNN converted image recognition neural network. Load reduction due to the application control is above 80%.

5) Joint optimization of connect-compute resources for edge learning and inference

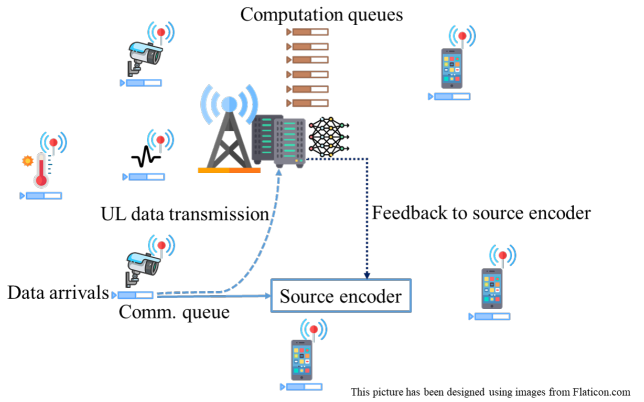


FIGURE 17. Adaptive data compression for energy-efficient edge inference

Reducing the communication cost and overhead can be done in several ways. Going beyond the previous section based on SNN architectures, this section proposes an alternative and complementary step that involves data compression before transmission, with target inference performance. Indeed, once the workload is placed across different nodes in the network (e.g., as proposed in previous sections), a joint allocation of wireless (e.g., precoding, decoding, transmit power, bandwidth) and computing (e.g., CPU scheduling) resources is fundamental to achieve the needed levels of energy efficiency, latency, and reliability. In addition, data compression and/or quantization determine learning and inference accuracy/confidence of ML models deployed at the edge of wireless networks, thanks to the MEC paradigm as described in Section IV-A1. The latter also calls for a paradigm shift from *data-oriented* to semantic and *goal-oriented* communications [12], [18], whose main objective is not to reliably transmit information, but to retrieve, at the receiver side, the relevant information needed to accomplish a task with target reliability (also known as goal-effectiveness [72]). Within this vision, this section focuses on the joint allocation of radio and computing resources to enable energy efficient, reliable, and timely edge inference. Therefore, performance indicators include energy consumption, delay, and inference confidence, translated also into inference accuracy.

The scenario, proposed in [13], [73], comprises multiple (possible heterogeneous) end devices, collecting data, compressing them, and uploading them to a Mobile Edge Host (MEH), through the wireless connection with an AP, as in Fig. 17. As shown in the figure, local communication and remote computation buffers model the E2E service delay. At the MEH, computing resources are shared among all users, which compete to access a sufficient pool guaranteeing their end-to-end delay requirement. In this type of connect-compute applications, the end-to-end delay comprises communication (to transmit data, mainly affecting uplink communications [74]) and computation delays (to process data). More specifically, we present numerical results obtained with the solution developed in [73]. The goal is to adaptively allocate communication resources (data compression and transmit power), and computation resources (local computing to compress data and remote CPU scheduling), to minimize the end devices sum energy consumption under two long-term constraints: *i)* end-to-end delay constraint, including local buffering, uplink transmission, remote buffering and processing; *ii)* average inference confidence, measured by the entropy at the output of a neural network classifying images [73]. The latter also translates into a correct classification rate. It is assumed that the MEH, based on measured levels of inference confidence, feeds the required compression scheme to the source encoder of the end devices. An example is shown in the figure with one devices, but the solution is applied to all devices.

In this specific example, we focus on an edge classification task on JPEG compressed images. To deal with dynamically evolving parameters (i.e., wireless channels and data arrivals) the system is organized in slots of equal duration. At the beginning of each slot, a decision is taken on: *i)* data compression (i.e., which JPEG compression level to select for transmitting images), *ii)* local computing resources to compress data, *iii)* uplink transmit power, and *iv)* MEH's CPU scheduling. Thanks to theoretical tools of Lyapunov stochastic network optimization [75], an online algorithm has been developed to jointly take these decisions in a per-slot basis, by only observing current wireless and computing resource conditions, as well as properly defined state variables that capture the behavior of the system in terms of congestion (i.e., communication and computation buffers state), and constraint violations (i.e., *virtual queues* that grow each time the inference confidence constraint is not satisfied, and are drained otherwise). More technical details can be found in [73]. The algorithm is tested on a similar scenario as the one presented in [73], with 6 devices, each one of them

requesting a different inference confidence level, comprising the two extreme benchmarks: *i)* the minimum energy device, i.e., the one transmitting the data with the maximum data compression, and *ii)* the maximum accuracy device, i.e., the one transmitting data with the maximum number of bits. Edge inference is performed with a pre-trained state of the art architecture, whose details are available in [73], assumed to be pre-uploaded at the MEH, on the CIFAR-10 data set [76]. A wireless AP is placed at the center of a circle of radius 100 m, and is equipped with an MEH with maximum CPU clock frequency 10 GHz, and the 6 devices uniformly randomly located inside the circle, transmitting at maximum power 20 dBm. Denoting by f_c the carrier frequency in GHz, and by d_k the distance in meters between device k and the AP, the channel gain is generated with path loss (in dB) $\mathrm{PL}_k = 33 + 25.50\log_{10}(d_k) + 20\log_{10}(f_c)$, and with Rayleigh fading with unit variance, changing across time slots, whose duration is set to 25 ms. In Fig. 18, numerical

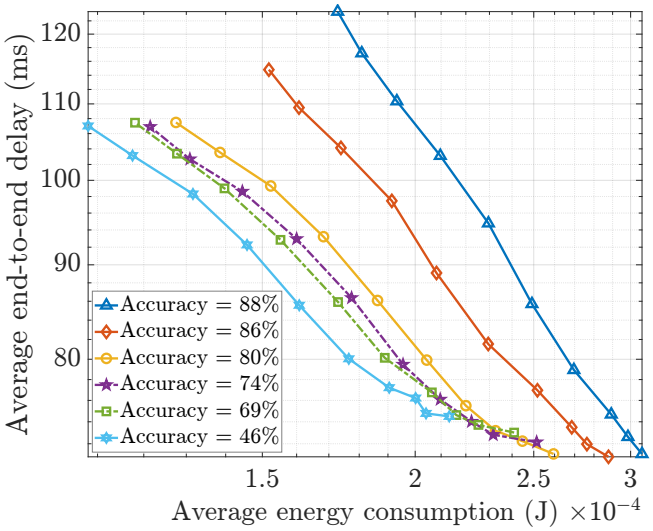


FIGURE 18. Trade-off between device energy consumption, end-to-end delay, and inference accuracy in computation offloading for edge inference services

results assessing the performance of the proposed method in terms of trade-off between energy, delay, and inference accuracy, are presented. Since each curve shows the average end-to-end delay as a function of the average energy consumption for a specific user, the trade-off with respect to inference accuracy can be appreciated through the different curves. In particular, let us focus on the first benchmark, i.e., the minimum energy device (light blue curve). This user exhibits the best trade-off between energy and latency, however experiencing highly degraded inference accuracy (around 46%). At the same time, the best accuracy case (blue curve - 88%) is paid by a higher energy consumption, for the same end-to-end delay, i.e., the worst trade-off between energy and delay. More interestingly, there are intermediate cases, obtained through the proposed adaptive compression strategy, which experience better energy-delay trade-off when compared to the best accuracy case, however without dramatically degrading the accuracy performance. As an example, the orange curve loses around 2% of accuracy. The choice of the target accuracy highly depends on the specific application, and it affects the energy-delay balance. What is more important, is the capability of the method to adapt to the application requirements to strike the best trade-off between energy and delay, by attaining the desired inference accuracy performance.

In this section, we presented one more brick towards the management of computing workloads (in this case for edge inference), with a joint approach that encompasses communication delay and energy consumption, to move data from their source to a remotely hosted ML model to run the inference task. Future research directions involve the full management of inference workloads, in a unified framework involving placement (Section IV-A2), balancing (Section IV-A3), and resource allocation, all under the AIaaS umbrella (Section IV-A1). Also, goal-oriented and semantic communications can be used as a tool to further improve the performance of edge inference [77].

B. TOWARDS TRUSTWORTHY, DISTRIBUTED AI

In previous sections, we discussed how future 6G networks are expected to pave the way to innovative services that will make massive use of AI and ML techniques, with new trade-offs involving energy, delay, accuracy, and complexity. However, the design of AI systems must also comply with additional requirements towards trustworthy AI, such as transparency of AI models, security of AI models, and privacy of data owners.

In Hexa-X, trustworthiness of AI and ML, involving security and privacy, has been identified as one of the pillars in 6G to ensure data protection and privacy, as well as model robustness. Indeed, it represents one of the identified KVIs, together with sustainability and inclusion. Since AI and ML will play a significant role in the development and operation of 6G networks, attacks on learning systems can impact any application that relies on these technologies. The focus of Section IV-B1 is on the vulnerabilities of AI-enabled systems against adversarial attacks on the use case of AI-driven power allocation in D-MIMO, as an additional challenge related to these novel settings also discussed in Section III-C, and also the privacy of FL.

Sections IV-B2 and IV-B3 address trustworthiness and transparency of AI models by deploying eXplainable Artificial Intelligence (XAI) methods. The goal of XAI is to investigate tools and techniques aimed at opening the so-called opaque (or black-box) models (e.g., DNNs) or at devising intrinsically interpretable and accurate models (e.g., rule based systems), thus producing details and reasons regarding the functioning of the model itself. In our solutions, we also combine FL and XAI: the acronym Fed -XAI [78] stands for federated learning of XAI models and is conceived to provide a leap forward toward trustworthy AI. The objective of Fed-XAI consists in devising methodological and technological solutions as follows: on one hand, to leverage the

FL approach for privacy preservation during collaboratively training of ML/AI models. On the other hand, to ensure an adequate degree of explainability of the AI-based systems. Notably, Fed-XAI can be regarded as an enabler for several families of use cases envisioned for 6G. As an example, it has recently been proposed as an enabling technology in 6G systems for an automated vehicle networking use case [79]. Inferencing accuracy represents the most relevant KPI, which must be pursued together with the KVI of explainability. Model complexity can be considered as a proxy for the interpretability level and may be associated with other XAI metrics (e.g., based on surveys) to evaluate explainability.

1) Challenges and enablers to achieve Trustworthy AI

In this first part, we demonstrate how susceptible AI systems are to adversarial attacks through the obtained results of applying evasion attacks against AI-driven power allocation in a D-MIMO network. We also show how the privacy of FL can be enhanced using blind signature scheme [80] and multi-hop communication [81].

In 6G networks, use cases related to usage of AI/ML in wireless tasks such as beamforming and power allocation in D-MIMO such as the one presented in Section III-C1, could be more adversary-sensitive. One such adversarial attacks can be evasion attacks where an attacker deliberately manipulates the input to the system in a way that causes the system to make incorrect or undesirable decisions. In [82], we simulate a successful adversarial evasion attack against AI-driven power allocation model in a distributed MIMO network where potential attack sources are illustrated in Fig. 19. The CDFs (cumulative distribution functions) of

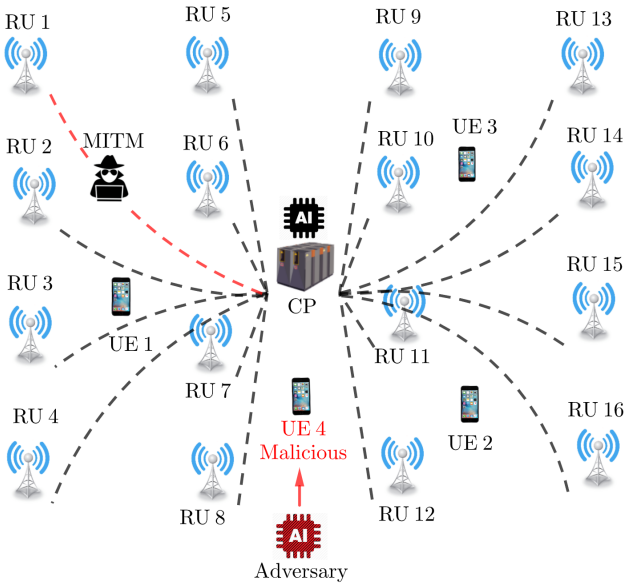


FIGURE 19. D-MIMO network with potential attacks.

per-user SE under different attack types are illustrated in Fig.20. The results show that our proposed attack has higher effects on degrading user spectral efficiencies in comparison to other conventional attacks (i.e., attacks where the attacker applies random perturbation such as gaussian noise). Also, we observe that the surrogate model (a model which is created by attacker by training a model on a similar or a subset of the input features of the target model) performance (marked as black-box) is marginally less disruptive than the original model performance (indicated as white-box). This finding shows that creating effective adversarial samples does not require the adversary to have access to the original AI model. As a result, it is necessary to adopt smart defense techniques to protect against such attacks. In the literature, potential countermeasures are proposed to

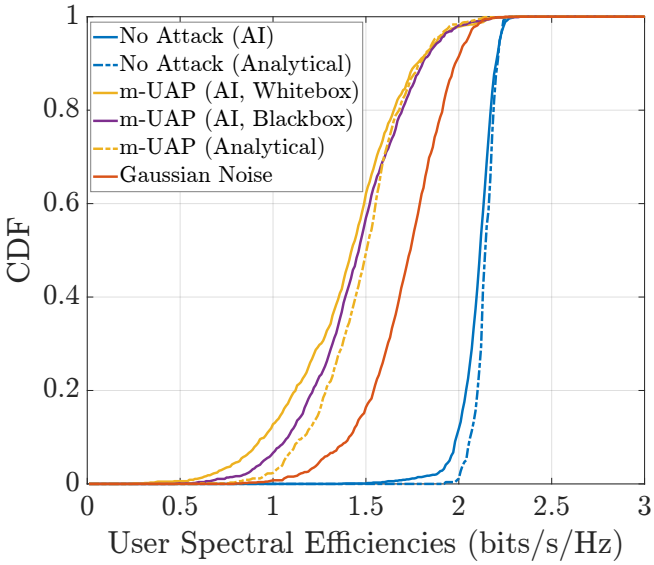


FIGURE 20. Comparison of different attack types ($\epsilon = 8$ dB).

create robust and resilient AI systems. Moving Target Defense (MTD) [83], which involves regularly updating the training data or model architecture, or introducing random perturbations to the model's parameters, is a technique used to increase the resiliency of AI systems. Input validation and robust learning approaches can be used to improve the robustness of ML models against poisoning attacks [84]. Defensive distillation [85] can also be used in combination with other techniques such as adversarial training to improve the robustness of machine learning models against evasion attacks. From privacy point of view, in 6G, large amount of user data will be used for training ML models. There are some privacy attacks such as membership inference attacks, model inversion attacks, and model extraction attacks that can be launched against machine learning models and systems to expose sensitive data. To protect sensitive data against this kind of attacks, privacy enhancing technologies (PETs) such as differential privacy [86], multi-party computation [87], and homomorphic encryption [88], also FL can be utilized. Federated learning, also introduced in Section IV-A3, is a collaborative machine learning techniques that preserves user data privacy by enabling users to learn a prediction ML

model in a collaborative way, while keeping the data on the individual devices. However, there are still security and privacy concerns in FL [89], [90], and it is a challenge to find solutions that provide both security and privacy at the same time. The solution, proposed in [13], [91] is a privacy solution that allows the run of security mechanisms to prevent model degradation, and take advantage of multi-hop communication and blind signature to preserve user privacy and prevent malicious behavior of clients. The example interactions between the server and clients are illustrated in Fig. 21. The proposed method can be regarded as an enabler for several use cases envisioned for 6G [27].

There are different metrics to measure the efficiency and effectiveness of the protective technical enablers for trustworthy AI. In term of privacy, the accuracy of the model and the overhead introduced by the privacy-preserving technique can be considered as relevant KPIs. In term of security, the adversarial attack success rate and adversarial defense success rate can be considered as relevant KPIs. Thus, by implementing right measures, organizations can help to reduce the risk of security and privacy attacks on their ML systems and protect the integrity and reliability of their ML models.

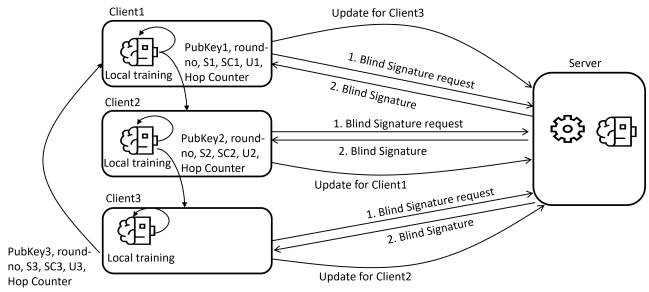


FIGURE 21. Example interactions between server and clients

2) Explainable AI for radio network control

Changing perspective towards XAI for automated radio network control, in this section we devise methods to explain and separate the effect of configuration and load to network performance KPIs. The KPI prediction task has been extensively studied in the literature [92], and is extremely relevant, e.g., whenever massive twinning and cobots applications as well as the resilient network infrastructure require high QoS [3]. The relevant measures of throughput, reliability, complexity reduction, accuracy, all important KPIs, can be predicted and problems can be anticipated and mitigated by ML methods. Another relevant measure is the accuracy of the XAI model compared to black box solutions, to improve the explanability while keeping the model accuracy unchanged or acceptably reduced.

The main technical difficulty in useable XAI for mobile radio KPI prediction is the complex causal relationship between configuration, load and performance. For example, antenna tilt configuration affects network performance indirectly via coverage, interference and cell load. These indirect

TABLE 2. Feature groups used in a sample mobile radio control dataset.

Name	Description
TA distribution	Timing Advance distribution of mobile terminals. It is derived from radio propagation delay measurements between the mobile terminal and base stations and can be used to estimate the distance from the base station. We normalize this distance with cell range, hence these features capture cell edge versus cell center distribution of mobile terminals.
Cell load	Various PM metrics describing user plane and signaling load of a given cell.
Interference	Measured uplink interference distribution in a given cell (downlink interference is unfortunately not available in our dataset).
Channel quality	Various downlink channel quality metrics including CQI and rank distributions.

metrics have a more direct effect on network performance, hence XAI will primarily find the importance of coverage, interference, and load and explain the predicted performance metric based on them, mostly ignoring the explanation of how network configuration affects this performance metric.

XAI models achieve flexibility and generalizability, as the explanation can be incorporated in arbitrary settings as expert knowledge. Model explanations enhance data quality as a pointwise explanation approach enables outlier and data error filtering by understanding model prediction and error. We also reach complexity gain, since the final model can be very simple by focusing on the key concepts rather than artifacts of the training data.

We proposed a first XAI model in [93] based on SHAP (SHapley Additive exPlanations) [94] and a Gradient Boosted Tree (GBT) regression model for KPI prediction. GBT models also enable deployment flexibility: tree-based regression models can be generated directly as program code without the need for ML or DL frameworks.

The main technical difficulty in giving the appropriate model explanation for network KPIs relies in the casual relation between network control, load and channel quality. SHAP explanation is calculated by considering subsets of all these measurable attributes and subtracting their contribution to the model. Since channel quality has the most direct effect on the network KPIs, the explanation will primarily find the importance of channel quality and explain the predicted KPI by putting lower importance on load and mostly ignoring the explanation of how control affects the KPI.

Reliably inferring causal relationships from observational data is generally considered to be impossible [95]. Rather than inferring the relationship, causal attributions [96], [97] assume to know the nature of the causal relationship based on domain knowledge and attempt to calculate attributions that respect these relationships. Asymmetric SHAP [96] is a method well suited for our task, since we are free to modify the order of the weights while computing SHAP for subsets, thus we can prioritize load over channel quality and control over all attributes.

We demonstrate our method on performance management

TABLE 3. Normalized average absolute feature attributions made by baseline and the best new method using two different causal ordering on four feature classes TA distribution (TA), cell load (Load), interference (I), and channel quality (CQ).

	TA	Load	I	CQ
TreeSHAP original	2.6%	34.9%	12.3%	50.2%
Load before TA	3.7%	40.1%	10.6%	45.6%
TA before Load	14.9%	33.4%	10.6%	41.1%

data from radio access network cells with 15 minutes granularity [93]. Model output is average downlink cell throughput for automatically determining the root cause of throughput degradations using explainers of the model. Input features of the model are described in Table 2. In Table 3, we compare the performance of the original TreeSHAP [98] method that does not take causal relations into account to two variants of our Asymmetric SHAP based method. In both methods, we consider Channel Quality as consequence of Interference, while Interference of TA distribution and Cell load. We have the option to order TA distribution before or after Cell load (or take the average of the two). Measurements indicate that we are able to move the assessed importance of the variable classes closer to the actual root cause.

3) Federated Learning of XAI models

The combination of FL paradigm and XAI techniques has recently gained increasing attention. Most existing solutions revolve around the original proposal of Federated Averaging (FedAvg) [99], as a method for executing Stochastic Gradient Descent (SGD) in a federated manner, and exploit post-hoc explainability techniques, such as feature relevance [100] or counterfactual explanations [101]. The FL of interpretable-by-design models, instead, may require the design of ad-hoc federation strategies, possibly different from the traditional FedAvg when the learning procedure is not based on the optimization of a global differentiable objective function. Among highly interpretable models, Takagi–Sugeno-Kang Fuzzy Rule-Based Systems (TSK-FRBS) [102] have been recently investigated and adapted for addressing regression tasks in a federated setting [103], [104]. We recall that a TSK-FRBS adopts linguistic if-then rules; an example of the generic k^{th} rule is reported in the following:

$$\begin{aligned} R_k : \textbf{IF}\ \ & X_1\ is\ A_{1,j_{k,1}}\ \ \textbf{AND}\ \ \ldots\ \ \textbf{AND}\ \ X_F\ is\ A_{F,j_{k,F}} \\ \textbf{THEN}\ \ & y_k = \gamma_{k,0} + \sum_{i=1}^{F} \gamma_{k,i} \cdot x_i \end{aligned} \tag{1}$$

where F is the total number of input variables, $A_{i,j_{k,i}}$ identifies the j^{th} fuzzy set of the fuzzy partition over the i^{th} variable considered in the k^{th} rule, and $\gamma_{k,i}$ are the coefficient of the linear model, with $i = 0, \ldots, F$.

We propose an approach for FL of TSK-FRBSs [105] from data, which is not iterative but rather generates the global model in one-shot: in a nutshell, first a local TSK-FRBS is generated by each participant and sent to the central server. The server is in charge of aggregating the received rule bases by juxtaposing them and resolving possible rule conflicts (i.e., rules with same antecendents and different consequents). Finally, the aggregated model is broadcast to the clients. Notably, a high level of interpretability is ensured thanks to the adoption of fuzzy uniform partitions with a limited number of fuzzy sets and an inference strategy based on the maximum voting (i.e., the output value depends only on a single rule, the one with the highest strength of activation) rather than the classical weighted averaging method, (i.e., the output value depends on all the rules activated by an input instance).

The experimental evaluation of the proposed FL strategy [105] is performed by comparing it with two alternative learning settings, namely centralized and local. In the former, local datasets are collected in a server for centralized processing. This setting represents the ideal case in which the entire dataset is available for model training, but evidently violates the requirement of data privacy. In the latter, each participant builds a model based on local data: the local approach guarantees privacy preservation but entails no form of collaboration among participants. The proposed FL of TSK-FRBSs is tested on four benchmark regression datasets, namely Weather Izmir (number of features $F = 9$, number of samples $N = 1461$), Treasury ($F = 15$, $N = 1049$), Mortgage ($F = 15$, $N = 1049$) and California ($F = 8$, $N = 20460$), considering 5 participants and using 5-fold cross-validation to assess model generalization capability. The distributed setting is simulated by randomly splitting each dataset in five chunks (one for each participant) with the same number of instances. Further details on the experimental setting are available in [105]. Fig. 22 summarizes the results evaluated in terms of MSE on the test sets. The height of the bars and the error bars represent the average value over the participants and the standard deviation, respectively. It can be observed that the FL scheme achieves better results, on average, compared to models generated locally: this outcome empirically demonstrates the benefit of participating in the FL process. The FL scheme approaches the centralized setting for Weather Izmir and California, while it is outperformed on Treasury and Mortgage, possibly due to the high dimensionality (15) and low overall number of samples (1049) of the two datasets.

Fed-XAI, and specifically FL of inherently interpretable models, helps improve the KPI of inferencing accuracy, still preserving the privacy of data owners and the KVI of explainability, which are regarded as pillars towards trustworthy distributed AI. The scope for further developments in the Fed-XAI area includes the design of ad-hoc federation strategies for other highly interpretable models (e.g., decision trees) and for challenging scenarios, i.e., when data are collected in streaming and possibly distributed in a non-i.i.d. manner among various clients.

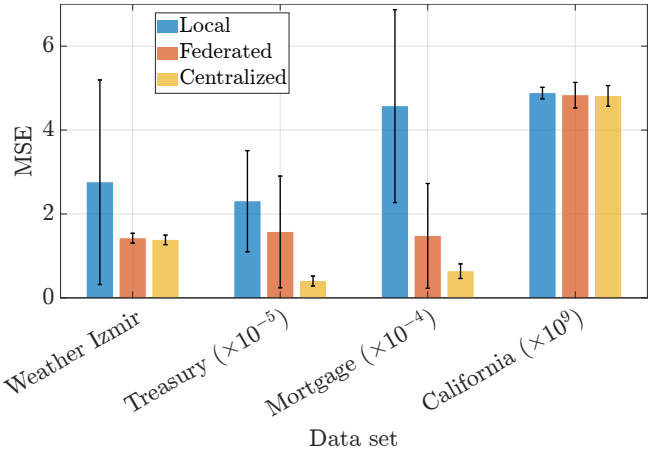


FIGURE 22. Experimental results: average MSE on four regression datasets. Comparison between local, federated and centralized learning schemes. Error bars represent standard deviation.

V. REGULATION ASPECTS

Several regulation and standardization bodies focus on AI and ML in future networks. Among the others, focusing on trustworthiness aspects, we focus on one specific action. In particular, considering the potential and future relevance of AI systems, the European Commission (EC) is currently in the process of developing an Artificial Intelligence regulation entitled AI Act [25]. We believe that the aspects described in the following are extremely relevant for future communication networks, in which AI will be a native component. The objectives of the AI Act are as follows:

- ensure that AI systems placed and used on the Union market and used are safe and respect existing law on fundamental rights and Union values;
- ensure legal certainty to facilitate investment and innovation in AI;
- enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems;
- facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

As AI components are expected to play a key role in next generation communication systems, it is important to understand the inherent AI Act requirements and to include corresponding solutions in the system design in order to maintain continued access to the European single market.

The AI Act differentiates various types of AI systems:

- Minimal or no risk AI applications: permitted without restrictions;
- Transparency risk AI applications (e.g., impersonation, bots): permitted but subject to information/transparency obligations;
- High risk AI applications (e.g., recruitment, medical devices, etc.): permitted subject to compliance with AI requirements and ex-ante conformity assessment;
- Unacceptable risk AI applications (e.g., social scoring): prohibited.

A list of systems considered to be High Risk is outlined in Annex III of the AI Act [25]. It is currently under debate in the EU Parliament and Council to which extent cellular systems fall into the High Risk category. Any High Risk system will be required to demonstrate compliance to a number of essential requirements in order to obtain access to the Singe European Market. Otherwise, market access will not be granted. The implementation of the regulation is - in simplified terms - relying on the following three key steps:

- The European Commission is publishing the regulation (the publication of the AI Act is expected for 2024).
- The European Commission is providing a Standardization Request (SR) to European Standardization Organizations (ESOs). An initial SR is expected for April 2023.
- ESOs finally build on the SR and develop Harmonised European Norms and other deliverables in support of the regulation.

All stakeholders, including industry, academia and others, are able to contribute to and influence in particular the 3rd step above through participation in the standardization process. A number of organizations are currently engaged in the development of standards of relevance to the AI Act: *i)* ISO/IEC JTC1 SC42 is developing international standards in the field of AI; *ii)* CEN/CENELEC are expected to adopt relevant ISO/IEC JTC SC42 specifications as European Norms such that they can be used for demonstrating compliance to the AI Act; *iii)* the Institute of Electrical and Electronics Engineers (IEEE) is developing global socio-technical standards in AI Ethics and Governance and is currently in discussion with ISO/IEC for international adoption of respective standards; *iv)* finally, ETSI has published a white paper [62] summarizing available deliverables which can be used to support the implementation of the AI Act; furthermore, the white paper outlines future plans of ETSI in the field of Human Factor, testing of AI systems, etc. In the European Standards Organizations, the specific detailed technical requirements and related testing procedures will be defined and will eventually be applied to determine product compliance and thus granting access to the European single market.

VI. CONCLUSIONS

We provided an overview of the Hexa-X activities around the topic of in-network AI and ML for 6G. We first introduced the UCs, KPIs and KVIs identified by the project, with special focus on those that have mostly affected the activities related to AI and ML-driven communication-computation co-design. Second, we presented a set of down selected technical enablers that we envision to enable the 6G ecosystem with the required performance. We focused on the two paradigm of *learning to communicate and communicating to learn*. Starting from the KPIs, the technical solutions are accompanied by quantifiable metrics that are evaluated through numerical simulations. These metrics include estimation errors, throughput, block error rate, capacity, beam scanning

time, load reduction, E2E delay and energy consumption, spectral efficiency, and attack success rate. All together, we believe them to enable 6G with challenging target values covering sustainability, trustworthiness, flexibility, and inclusion. Finally, we discussed part of the ongoing regulation activities related to AI, along with their impact in future communication networks and research.

A lot of work still needs to be done around the topics presented in this paper and beyond, towards the 6G standardization efforts that should produce first outputs around 2030. Overall, AI and ML will be native components of future communication networks, pushing researchers to identify new key challenges and technical enablers to enhance performance toward new unexplored limits, while not forgetting about fundamental values that include sustainability, trustworthiness, and inclusion.

REFERENCES

[1] The 5G Infrastructure Association, "European vision for the 6G network ecosystem," 2021. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2021/06/WhitePaper-6G-Europe.pdf

[2] 6G Smart Networks and Services Industry Association, "What societal values will 6G address? ," May 2022. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2022/05/What-societal-values-will-6G-address-White-Paper-v1.0-final.pdf

[3] Nurul H. Mahmood *et al.*, "White paper on critical and massive machine type communication towards 6G," June 2020. [Online]. Available: http://jultika.oulu.fi/files/isbn9789526226781.pdf

[4] E. Calvanese Strinati *et al.*, "6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 42–50, 2019.

[5] Huawei, "6G: The Next Horizon White Paper," January 2022. [Online]. Available: https://www.huawei.com/en/technology-insights/future-technologies/6g-white-paper

[6] Samsung, "The Next Hyper Connected Experience for All," 2020. [Online]. Available: https://cdn.codeground.org/nsr/downloads/researchareas/20201201_6G_Vision_web.pdf

[7] Nokia, "Technology innovations for 6G system architecture," April 2022. [Online]. Available: https://www.bell-labs.com/institute/white-papers/technology-innovations-for-6g-system-architecture/

[8] Ericsson, "Connecting a cyber-physical world," February 2022. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/a-research-outlook-towards-6g

[9] NTT DOCOMO, INC., "5G Evolution and 6G ," January 2022. [Online]. Available: https://www.docomo.ne.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v4.0.pdf

[10] Orange, "White Paper: Orange's vision for 6G," March 2022. [Online]. Available: https://hellofuture.orange.com/en/

[11] M. A. Uusitalo *et al.*, "6g vision, value, use cases and technologies from european 6g flagship project hexa-x," *IEEE Access*, vol. 9, pp. 160 004–160 020, 2021.

[12] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6g: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2022.

[13] "Hexa-X Deliverable D4.2 - AI-driven communication & computation co-design: initial solutions," 2022. [Online]. Available: https://hexa-x.eu/wp-content/uploads/2022/07/Hexa-X_D4.2_v1.0.pdf

[14] "Hexa-X Deliverable D4.3 - AI-driven communication & computation co-design solutions," 2023. [Online]. Available: https://hexa-x.eu/wp-content/uploads/2023/05/Hexa-X_D4.3_v1.0.pdf

[15] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[16] E. Peltonen *et al.*, "6g white paper on edge intelligence," 2020. [Online]. Available: https://arxiv.org/abs/2004.14850

[17] N. Kato, B. Mao, F. Tang, Y. Kawamoto, and J. Liu, "Ten challenges in advancing machine learning technologies toward 6G," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 96–103, 2020.

[18] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[19] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 957–975, 2020.

[20] S. Zhang and D. Zhu, "Towards artificial intelligence enabled 6G: State of the art, challenges, and opportunities," *Computer Networks*, vol. 183, p. 107556, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138912862031207X

[21] J. Wu *et al.*, "Toward native artificial intelligence in 6G networks: System design, architectures, and paradigms," 2021. [Online]. Available: https://arxiv.org/abs/2103.02823

[22] M. K. Shehzad, L. Rose, M. M. Butt, I. Z. Kovács, M. Assaad, and M. Guizani, "Artificial intelligence for 6g networks: Technology advancement and standardization," *IEEE Vehicular Technology Magazine*, vol. 17, no. 3, pp. 16–25, 2022.

[23] W. Guo, "Explainable artificial intelligence for 6G: Improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[24] M. C. Filippou *et al.*, "Pervasive Artificial Intelligence in Next Generation Wireless: The Hexa-X Project Perspective," 2022. [Online]. Available: https://ceur-ws.org/Vol-3189/paper_05.pdf

[25] "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[26] "Hexa-X Deliverable D1.2 - Expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum," 2021. [Online]. Available: https://hexa-x.eu/wp-content/uploads/2021/05/Hexa-X_D1.2.pdf

[27] "Hexa-X Deliverable D1.3 - Targets and requirements for 6G - initial E2E architecture," 2022. [Online]. Available: https://hexa-x.eu/wp-content/uploads/2022/03/Hexa-X_D1.3.pdf

[28] D. Neumann, T. Wiese, and W. Utschick, "Learning the mmse channel estimator," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2905–2917, 2018.

[29] Y. Chen, J. Mohammadi, S. Wesemann, and T. Wild, "Turbo-AI, Part I: Iterative machine learning based channel estimation for 2D massive arrays," in *in Proc. IEEE 93rd Veh. Technol. Conf. (VTC'21 Spring), Apr. 2021.*, 2021, pp. 1–6.

[30] ——, "Turbo-AI, Part II: Multi-dimensional iterative ml-based channel estimation for B5G," in *in Proc. IEEE 93rd Veh. Technol. Conf. (VTC'21 Spring), Apr. 2021.*, 2021, pp. 1–6.

[31] K. Yu *et al.*, "Second order statistics of nlos indoor mimo channels based on 5.2 GHz measurements," in *GLOBECOM'01. IEEE Global Telecommunications Conference (Cat. No.01CH37270)*, vol. 1, 2001, pp. 156–160 vol.1.

[32] T. Yassine and L. Le Magoarou, "mpNet: variable depth unfolded neural network for massive MIMO channel estimation," *IEEE Trans. Wireless Commun.*, vol. PP, pp. 1–1, 2022.

[33] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *arXiv preprint arXiv:2012.08405*, Dec. 2020.

[34] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[35] B. Chatelier, L. Le Magoarou, and G. Redieteab, "Efficient deep unfolding for siso-ofdm channel estimation," Oct. 2022.

[36] J. M. Mateos-Ramos, C. Häger, M. F. Keskin, L. L. Magoarou, and H. Wymeersch, "Model-driven end-to-end learning for integrated sensing and communication," Dec. 2022.

[37] D. L. Dampahalage, K. B. S. Manosha, N. Rajatheva, and M. Latva-Aho, "Weighted-Sum-Rate Maximization for an Reconfigurable Intelligent Surface Aided Vehicular Network," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 687–703, 2021.

[38] Z. Zhang and L. Dai, "Capacity Improvement in Wideband Reconfigurable Intelligent Surface-Aided Cell-Free Network," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[39] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent Reflecting Surface Aided MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Journal on Selected Areas in Comm.*, vol. 38, no. 8, pp. 1719–1734, 2020.

[40] E. Calvanese Strinati *et al.*, "Reconfigurable, intelligent, and sustainable wireless environments for 6G smart connectivity," *IEEE Communications Magazine*, vol. 59, no. 10, pp. 99–105, 2021.

[41] D. Dampahalage, K. B. Shashika Manosha, N. Rajatheva, and M. Latva-Aho, "Supervised Learning Based Sparse Channel Estimation For RIS Aided Communications," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8827–8831.

[42] J. A. Tropp and A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[43] H. Farhadi and M. Sundberg, "Machine learning empowered context-aware receiver for high-band transmission," in *2020 IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1–6.

[44] H. Farhadi, J. Haraldsson, and M. Sundberg, "A deep learning receiver for non-linear transmitter," *IEEE Access*, 2023.

[45] D. Korpi, M. Honkala, J. Huttunen, F. A. Aoudia, and J. Hoydis, "Waveform learning for reduced out-of-band emissions under a nonlinear power amplifier," 2021. [Online]. Available: https://arxiv.org/abs/2201.05524

[46] J. Pihlajasalo, D. Korpi, M. Honkala, J. M. Huttunen, T. Riihonen, J. Talvitie, A. Brihuega, M. A. Uusitalo, and M. Valkama, "HybridDeepRx: Deep learning receiver for high-EVM signals," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2021.

[47] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 1251–1275, 2020.

[48] Y. Zhao, I. G. Niemegeers, and S. H. De Groot, "Power allocation in cell-free massive MIMO: A deep learning method," *IEEE Access*, vol. 8, pp. 87 185–87 200, 2020.

[49] C. D'Andrea, A. Zappone, S. Buzzi, and M. Debbah, "Uplink power control in cell-free massive MIMO via deep learning," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 554–558.

[50] T. Van Chien, T. Nguyen Canh, E. Björnson, and E. G. Larsson, "Power control in cellular massive MIMO with varying user activity: A deep learning solution," *IEEE Transactions on Wireless Communications*, vol. 19, no. 9, pp. 5732–5748, 2020.

[51] N. Rajapaksha, K. B. S. Manosha, N. Rajatheva, and M. Latva-aho, "Deep learning-based power control for cell-free massive MIMO networks," in *ICC 2021 - 2021 IEEE International Conference on Communications (ICC)*, 2021.

[52] H. Masoumi and M. J. Emadi, "Performance analysis of cell-free massive MIMO system with limited fronthaul capacity and hardware impairments," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1038–1053, 2020.

[53] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max–min SINR of cell-free massive MIMO systems," *IEEE Transactions on Wireless Comm.*, vol. 18, no. 4, pp. 2021–2036, 2019.

[54] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," 2010, pp. 399–406.

[55] X. Chen, J. Liu, Z. Wang, and W. Yin, "Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds," in *Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, Canada*, 2018, pp. 9079–9089.

[56] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. of Information Theory and Applications Workshop (ITA)*, Feb 2019, pp. 1–8.

[57] S. W. Andrzej Goscinski, Elisa Bertino, "Guest editor's introduction: Special section on edge AI as a service," *IEEE Transactions on Services Computing*, vol. 15, no. 2, 2022.

[58] W. L. Wenyu Zhang, Sherali Zeadally *et al.*, "Edge ai as a service: Configurable model deployment and delay-energy optimization with result quality constraints," *IEEE Transactions on Cloud Computing*, 2022.

[59] A. M. Valerio De Caro, Saira Bano *et al.*, "AI-as-a-service toolkit for human-centered intelligence in autonomous driving," *IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2022.

[60] L. Xuncheng, W. Jingyi, Z. Weizhan, and Z. Qinghua, "Mobile real-time facial expression tracking with the assistant of public ai-as-a-service," *IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2020.

[61] M. D. Mueck, A. E. B. On, and S. D. Boispean, "Upcoming european regulations on artificial intelligence and cybersecurity," *to appear in IEEE Communicatons Magazine*, 2023.

[62] M. D. Mueck *et al.*, "ETSI activities in the field of artificial intelligence - preparing the implementation of the european ai act," *ETSI White Paper*, vol. 52, 2022.

[63] A. Mazloomi, H. Sami, J. Bentahar, H. Otrok, and A. Mourad, "Reinforcement learning framework for server placement and workload allocation in multi-access edge computing," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[64] S. Tuli, G. Casale, and N. R. Jennings, "SplitPlace: AI augmented splitting and placement of large-scale neural networks in mobile edge environments," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2022.

[65] S. Demirci, S. Sagiroglu, and M. Demirci, "Energy-efficient virtual security function placement in nfv-enabled networks," *Sustainable Computing: Informatics and Systems*, vol. 30, p. 100494, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210537920302171

[66] L. Magoula, S. Barmpounakis, I. Stavrakakis, and N. Alonistioti, "A genetic algorithm approach for service function chain placement in 5G and beyond, virtualized edge networks," *Computer Networks*, vol. 195, p. 108157, May 2021.

[67] N. Alharbe, A. Aljohani, and M. A. Rakrouki, "A fuzzy grouping genetic algorithm for solving a real-world virtual machine placement problem in a healthcare-cloud," *Algorithms*, vol. 15, no. 4, 2022. [Online]. Available: https://www.mdpi.com/1999-4893/15/4/128

[68] Z. Michalewicz and M. Schoenauer, "Schoenauer, m.: Evolutionary algorithms for constrained parameter optimization problems. evolutionary computation 4(1), 1-32," *Evolutionary Comp.*, vol. 4, pp. 1–32, March 1996.

[69] S. Mitchell, M. J. O'Sullivan, and I. Dunning, "PuLP : A Linear Programming Toolkit for Python," 2011.

[70] Z. Zvara, P. G. Szabó, B. B. Lóránt, and A. A. Benczúr, "System-aware dynamic partitioning for batch and streaming workloads," in *Proceedings of the 14th IEEE/ACM International Conference on Utility and Cloud Computing*, 2021, pp. 1–10.

[71] N. Rathi, G. Srinivasan, P. Panda, and K. Roy, "Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[72] M. Merluzzi, M. C. Filippou, L. G. Baltar, and E. C. Strinati, "Effective goal-oriented 6G communications: the energy-aware edge inferencing case," in *2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2022, pp. 457–462.

[73] M. Merluzzi, C. Battiloro, P. Di Lorenzo, and E. C. Strinati, "Energy-efficient classification at the wireless edge with reliability guarantees," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2022, pp. 109–114.

[74] J. Oueis and E. Calvanese Strinati, "Uplink traffic in future mobile networks: Pulling the alarm," in *Proc. of Int. Conf. on Cognitive Radio Oriented Wireless Networks, Springer*, mai 2016.

[75] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool Publ., 2010.

[76] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009.

[77] M. Merluzzi, M. C. Filippou, L. G. Baltar, M. D. Muek, and E. C. Strinati, "6G goal-oriented communications: How to coexist with legacy systems?" 4 2023. [Online]. Available: https://www.techrxiv.org/articles/preprint/6G_goal-oriented_communications_How_to_coexist_with_legacy_systems_/22189879

[78] J. L. C. Bárcena *et al.*, "Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models," in *3rd Italian Workshop on Explainable Artificiale Intelligence, XAI.it 2022*, vol. 3277, 2022. [Online]. Available: https://ceur-ws.org/Vol-3277/paper8.pdf

[79] A. Renda *et al.*, "Federated Learning of Explainable AI Models in 6G Systems: Towards Secure and Automated Vehicle Networking," *Information*, vol. 13, no. 8, p. 395, 2022.

[80] D. Chaum, "Blind signatures for untraceable payments," in *Advances in Cryptology: Proceedings of CRYPTO '82, Santa Barbara, California, USA, August 23-25, 1982*, D. Chaum, R. L. Rivest, and A. T. Sherman, Eds. Plenum Press, New York, 1982, pp. 199–203.
[81] A. Blanco-Justicia, J. Domingo-Ferrer, S. Martínez, D. Sánchez, A. Flanagan, and K. E. Tan, "Achieving security and privacy in federated learning systems: Survey, research challenges and future directions," *Eng. Appl. Artif. Intell.*, vol. 106, p. 104468, 2021.
[82] Ö. Faruk Tuna, F. E. Kadan, and L. Karaçay, "Practical adversarial attacks against ai-driven power allocation in a distributed mimo network," *arXiv e-prints*, pp. arXiv–2301, 2023.
[83] S. Sengupta, T. Chakraborti, and S. Kambhampati, "Mtdeep: boosting the security of deep neural nets against adversarial attacks with moving target defense," in *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.
[84] Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "Ai and 6g security: Opportunities and challenges," in *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2021, pp. 616–621.
[85] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 582–597.
[86] O. Haliloglu, E. U. Soykan, and A. Alabbasi, "Privacy preserving federated rsrp estimation for future mobile networks," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.
[87] O. Goldreich, "Secure multi-party computation," *Manuscript. Preliminary version*, vol. 78, no. 110, 1998.
[88] X. Yi, R. Paulet, E. Bertino, X. Yi, R. Paulet, and E. Bertino, *Homomorphic encryption*. Springer, 2014.
[89] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.
[90] E. U. Soykan, L. Karaçay, F. Karakoç, and E. Tomur, "A survey and guideline on privacy enhancing technologies for collaborative machine learning," *IEEE Access*, vol. 10, pp. 97 495–97 519, 2022.
[91] F. Karakoç *et al.*, "A security-friendly privacy solution for federated learning," in *AI6G workshop, IEEE WCCCI*, 2022.
[92] J. L. C. Bárcena *et al.*, "Towards trustworthy ai for QoE prediction in B5G/6G networks," in *AI6G workshop, IEEE WCCCI. CEUR WORKSHOP PROCEEDINGS*, vol. 3189, 2022.
[93] D. M. Kelen, P. Kersch, and A. Benczúr, "Causal explanations for performance in radio networks," in *AI6G workshop, IEEE WCCCI. CEUR WORKSHOP PROCEEDINGS*, vol. 3189, 2022.
[94] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural inf. proc. systems*, vol. 30, 2017.
[95] C. Winship and S. L. Morgan, "The estimation of causal effects from observational data," *Annual review of sociology*, pp. 659–706, 1999.
[96] C. Frye, C. Rowat, and I. Feige, "Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1229–1239, 2020.
[97] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models," *Advances in neural information processing systems*, vol. 33, pp. 4778–4789, 2020.
[98] I. Covert, S. M. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation." *J. Mach. Learn. Res.*, vol. 22, pp. 209–1, 2021.
[99] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.
[100] J. Fiosina, "Interpretable privacy-preserving collaborative deep learning for taxi trip duration forecasting," in *International Conference on Vehicle Technology and Intelligent Transport Systems, International Conference on Smart Cities and Green ICT Systems*. Springer, 2022, pp. 392–411.
[101] P. Chen, X. Du, Z. Lu, J. Wu, and P. C. Hung, "EVFL: An explainable vertical federated learning for data-oriented artificial intelligence systems," *Journal of Systems Architecture*, vol. 126, p. 102474, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1383762122000583
[102] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 1, pp. 116–132, Jan 1985.
[103] A. Wilbik and P. Grefen, "Towards a federated fuzzy learning system," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.
[104] X. Zhu, D. Wang, W. Pedrycz, and Z. Li, "Horizontal federated learning of takagi-sugeno fuzzy rule-based models," *IEEE Transactions on Fuzzy Systems*, 2021.
[105] J. L. C. Bárcena, P. Ducange, A. Ercolani, F. Marcelloni, and A. Renda, "An Approach to Federated Learning of Explainable Fuzzy Regression Models," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2022, pp. 1–8.

MATTIA MERLUZZI (Member, IEEE) received the M.S. degree in Telecommunication Engineering and the Ph.D. degree in Information and Communication Technologies from Sapienza University of Rome, Italy, in 2017 and 2021, respectively. He is currently a research scientist at CEA-Leti, Grenoble, France, where he is involved in the research team of the H2020 project Hexa-X. He has participated in the H2020 EU/Japan project 5G-Miedge and the H2020 EU/Taiwan project 5G CONNI. His primary research interests are in edge computing, beyond 5G systems, stochastic optimization, and edge machine learning. He was the recipient of the 2021 GTTI (Italian National Group on Telecommunications and Information Theory) Award for the Best Ph.D. thesis.

TAMÁS BORSOS received his M.Sc. in computer sciences in 1998 from the Budapest University of Technology and Economics in Hungary. He joined Ericsson Research and has been driving concept development in the area of 3G and LTE network management, real-time analytics and in innovation project on precise localization. He is currently senior specialist in research area Artificial Intelligence, focusing on AI integration in B5G/6G networks.

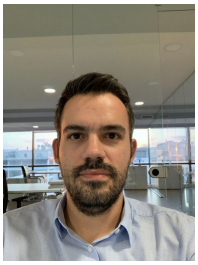
NANDANA RAJATHEVA (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 1987, and the M.Sc. and Ph.D. degrees from the University of Manitoba, Winnipeg, MB, Canada, in 1991 and 1995, respectively. He is currently a Professor with the Centre for Wireless Communications, University of Oulu, Finland. During his graduate studies, he was a Canadian Commonwealth Scholar in Manitoba. From 1995 to 2010, he was a Professor with the University of Moratuwa and the Asian Institute of Technology, Thailand. He is currently leading the AI-driven air interface design task in Hexa-X EU Project and also active in Hexa-X II. He has coauthored more than 200 refereed papers published in journals and in conference proceedings. His research interests include physical layer in 6G, machine learning for PHY and MAC, integrated sensing and communications, and channel coding.

ANDRÁS BENCZÚR received his Ph.D. in applied mathematics at the Massachusetts Institute of Technology in 1997. Since then, he is affiliated with the Institute of Computer Science and Control in Hungary, where he is leading a research lab with focus on Data Science and AI and acting as principal investigator in several EU, national, and industrial projects. He serves on the program committees of leading conferences including WWW, WSDM, SIGKDD, SIGIR. He is scientific director of the Artificial Intelligence National Laboratory Hungary, a consortium of 11 institutions with over 200 researchers, and leads the Data Industry Working Group of the Hungarian Artificial Intelligence Coalition.

HAMED FARHADI is a Senior Researcher at Ericsson Research, Stockholm, Sweden. He received his PhD degree in Telecommunication from KTH Royal Institute of Technology, Stockholm, Sweden in 2014. He was a Postdoctoral Research Fellow at Harvard University, Cambridge, MA, USA in 2016, and a Researcher at Chalmers University of Technology, Gothenburg, Sweden in 2015. His research mainly lies in statistical signal processing and machine learning for wireless communication networks. Hamed is the editor of a book on 'Machine Learning - Advanced Techniques and Emerging Applications', and a book on 'Medical Internet of Things (m-IoT) - Enabling Technologies and Emerging Applications'. Hamed has been the recipient of IEEE ICASSP best student paper award. He has been on the editorial board of the Springer International Journal of Wireless Information Networks since 2015. He is the Technical Manager of the European 6G Flagship Project Hexa-X.

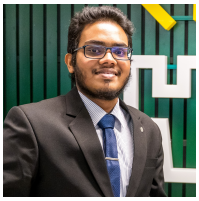
TAHA YASSINE received the engineering degree in computer science from the National Institute of Applied Sciences (INSA Rennes), France, and the M.Sc. degree in research in computer science from the University of Rennes 1, Rennes, France, in 2020. He is currently pursuing the Ph.D. degree with INSA Rennes, IETR, Rennes, and b<>com, Rennes. His current research topics include signal processing, wireless communications, and machine learning.

MARKUS MUECK received the Dipl.-Ing. and ing. dipl. degrees from the University of Stuttgart, Germany and the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France respectively in 1999. In 2006, he received the Doctorate degree of ENST in Communications. Dr. Mueck is a Principal Engineer with Intel Deutschland, Munich, Germany and an Engineering Director leading Intel Germany's Standardization organization; he is deeply involved in Artificial Intelligence (AI) standardization and European Policy related activities in the context of the European AI Act. As Chair of the ETSI OCG AI Committee overseeing ETSI's AI/ML activities, Dr. Mueck leads ETSI's delegation in discussions with the European Commission, CEN/CENELEC and other organizations. Furthermore, he acts as Vice-Chair of the ETSI Board, Member of the Board of the 5G Automotive Association (5GAA), he is member of the ETSI Delegation to 3GPP PCG/OP and Adj. Professor of University of Technology, Sydney, Australia. He is leading the Hexa-X (Europe's 6G Flagship project) Working Package on Artificial Intelligence and Machine Learning.



SOKRATIS BARMPOUNAKIS is a Senior Research Engineer and Solutions Architect for WINGS ICT Solutions, Athens Greece. He holds a Ph.D. in "Context-based Resource Management and Slicing for SDN-enabled 5G Smart, Connected Environments", since May 2018, from the National and Kapodistrian University of Athens. He obtained his Engineering Diploma in Electrical and Computer Engineering, from the National Technical University of Athens (NTUA). He has participated with technical and managerial roles in more than 20 European Projects and bilateral industrial contracts. His main fields of interest are: AI-native Beyond 5G and 6G Networks, Optimisation for B5G/6G networks, Energy/Trust/Sustainability aspects for B5G/6G Networks, Network architectures and protocols. He serves as a Technical Manager, WP/task leader for various projects, including the flagship EU project for 6G, Hexa-X-II. He is a member of the technical chamber of Greece.

EMILIO CALVANESE STRINATI (Master 2001, Ph.D 2005) worked at Motorola Labs between 2002 and 2006. In 2006 he joint CEA LETI. From 2010 to 2012, he has been the co-chair of the wireless working group in GreenTouch Initiative (on the future energy efficient communication networks). From 2011 to 2016 he was CEA's the Smart Devices & Telecommunications strategic programs Director, then the Smart Devices & Telecommunications Scientific and Innovation Director until 2020 . Since 2020 he is the Nanotechnologies and Wireless for 6G (New-6G) Program Director focusing on future 6G technologies. He has published around 200 papers in journals, international conferences, and books chapters, and he has given more than 200 international invited talks, keynotes and tutorials. He is the main inventor of more than 80 patents. His current research interests are on Reconfigurable Intelligent Surfaces, AI, Semantic-Goal-oriented communications in the contest of 6G.

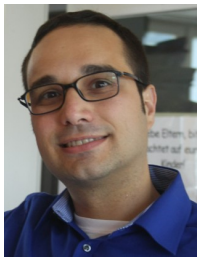


DILIN DAMPAHALAGE received B.Sc. (Hons.) degree in electronics and telecommunication engineering from University of Moratuwa, Sri Lanka, in 2018, and M.Sc. (Technology) degree in wireless communications engineering from University of Oulu, Finland, in 2020. He is currently a doctoral researcher at the Centre for Wireless Communications, University of Oulu, Finland. From May - August 2022, he completed a summer internship at Nokia Standards, Oulu, Finland. He is currently working under the 6G Flagship Program in University of Oulu, and actively involved in the AI-driven air interface design task in Hexa-X EU Project and its extension Hexa-X II. His research interests include reconfigurable intelligent surfaces, vehicular communications, channel estimation, and application of machine learning and optimization techniques in wireless communications.

PANAGIOTIS DEMESTICHAS (Senior Member IEEE) is a Professor at the University of Piraeus, School of ICT, Department of Digital Systems, Greece. Currently, he focuses on the development of systems for WINGS ICT Solutions (www.wings-ict-solutions.eu) and its spin-out Incelligent (www.incelligent.net). WINGS focuses on advanced solutions, leveraging on IoT / 5G / AI / AR, for the environment (air quality), utilities and infrastructures (water, energy, gas, transportation, construction), production and manufacturing (aquaculture, agriculture and food safety, logistics and industry 4.0), service sectors (health, security). Incelligent focuses on products for banking, the public sector and for telecommunication infrastructures. Panagiotis conducts research on 6G, cloud and IoT, big data and artificial intelligence, orchestration / diagnostics and intent-oriented mechanisms. He holds a Diploma and a Ph.D. degree on Electrical Engineering from the National Technical University of Athens (NTUA). He holds patents, has published numerous articles and research papers, and is a member of the Association for Computing Machinery (ACM).

PIETRO DUCANGE received the M.Sc. degree in Computer Engineering and the Ph.D. degree in Information Engineering from the University of Pisa in 2005 and 2009, respectively. Currently, he is an associate professor of Information Systems and Technologies at the University of Pisa, Italy. His main research interests include explainable artificial intelligence, big data mining, social sensing and sentiment analysis. He has been involved in a number of R&D projects in which data mining and computation intelligence algorithms have been successfully employed. He has co-authored over 100 papers in international journals and conference proceedings. He is a member of the Editorial Board for Soft Computing Journal.

MILTIADIS C. FILIPPOU (Senior Member, IEEE) was born in Athens, Greece, in 1984. He received the Dipl.Eng. degree in electrical and computer engineering from the National Technical University of Athens (NTUA), Greece, in 2007, and the Ph.D. degree in electronics and telecommunications from Telecom ParisTech, France, in 2014. From 2014 to 2015, he was a Research Fellow at the Institute for Digital Communications (IDCOM), The University of Edinburgh, U.K. Between 2015 and 2022, he was a Senior Standards and Research Engineer at Intel Germany, where he had received multiple patent and division awards. He is now a Lead Researcher with Nokia Germany. Between 2017 and 2022, he had served as a Delegate and Work Item Rapporteur for ETSI ISG MEC, where he had (co)-sourced more than 120 technical contributions accepted by the standard. Between 2021 and 2022, he led the AI/ML work of the EU 6G Flagship Project Hexa-X. He has authored and co-authored four book chapters and more than 50 technical papers, appearing in high-impact IEEE journals and major conference proceedings and has co-invented 10 granted patents. His current research interests include: AI/ML in network automation, task-oriented communications and edge computing.



LEONARDO GOMES BALTAR received his B.Sc. and M.Sc. degrees in Electrical Engineering from the Federal University of Rio de Janeiro (UFRJ), Brazil, and his Dr.-Ing. from the Technical University of Munich (TUM), Germany. From 2006 to 2015, he was with the Chair for Circuit Theory and Signal Processing at TUM as a research and teaching associate. His research was on digital signal processing for communications. He is currently a Senior Standards and Research Engineer at Intel Germany and an active contributor of the 5G Automotive Association, where he also chairs a WG, and of ETSI ITS.

JOHAN HARALDSON received the M.Sc. degree in computer science from Linköping University, Sweden, in 2007. He joined Ericsson the same year, where he has worked in several areas: including software development, technical sales support, and research. In 2014 he joined the AI Research department within Ericsson Research, and he is currently a Senior Researcher. His current interest is in applying machine learning to the physical layer.

LEYLI KARAÇAY received her Ph.D. and M.Sc. degrees both in Computer Science and Engineering in 2020 and 2012 from Sabanci University, Turkey. She had worked as a Teaching Assistant at Sabanci University for 7 years. She joined Ericsson Research Turkey in 2019 as a Security Researcher. Her research interests are privacy-enhancing technologies, ML/AI security, 5G and 5GB security.

DANI KORPI received the M.Sc. and D.Sc. degrees (Hons.) in communications engineering and electrical engineering from the Tampere University of Technology (TUT), Finland, in 2014 and 2017, respectively. He is currently a Senior Specialist with Nokia Bell Labs, Espoo, Finland. His Ph.D. thesis received the Best Dissertation of the Year Award from TUT, as well as the Finnish Technical Sector's Award for the best doctoral dissertation in 2017. His current research focuses on applying machine learning to wireless communications, especially on the physical layer. In particular, he is working on building a native foundation for machine learning in 6G radio networks.

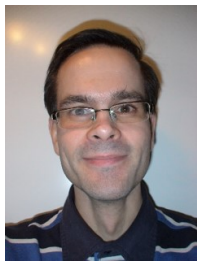
VASILIKI LAMPROUSI received the B.Sc. degree in Physics, and the M.Sc. in Materials Physics from National and Kapodistrian University of Athens, Greece, in 2011 and 2014, and she, also received the M.Sc. degree in Applied Data Science from the University of Gothenburg, Sweden, in 2020. Currently, she is a solution engineer at WINGS ICT Solutions. She has participated in various European Projects, among which Hexa-X, Hexa-X II and Dedicat 6G. Her research interests include optimization for B5G/6G networks and forefront development in Machine Learning, and Data Science.

FRANCESCO MARCELLONI (Member, IEEE) is full professor of Data Mining and Machine Learning at the University of Pisa. His main research interests include explainable artificial intelligence, federated learning, data mining for big data and streaming data, sentiment analysis and opinion mining, genetic fuzzy systems, and fuzzy clustering algorithms. He has co-edited three volumes, four journal special issues, and is (co-)author of a book and of more than 240 papers in international journals, books and conference proceedings. Recently, he has received the 2021 IEEE Transactions on Fuzzy Systems Outstanding Paper award and the 2022 IEEE Computational Intelligence Magazine Outstanding Paper award. He serves as associate editor of IEEE Transactions on Fuzzy Systems (IEEE), Information Sciences (Elsevier), Soft Computing (Springer), and is on the editorial board of a number of other international journals. He has coordinated various research projects funded by both public and private entities. He has also coordinated two Erasmus+ KA2 projects.

JAFAR MOHAMMADI (Member, IEEE) received the Ph.D. degree from the Technical University of Berlin, Berlin, Germany in 2016, on the topic of information-theoretic secrecy. He was a Researcher with the Heinrich Hertz Institute Fraunhofer from 2011 to 2017. In 2017, he joined Rutgers University, Piscataway, NJ, USA, as a Post-Doctoral Researcher working on differential privacy for distributed machine learning. He is currently a Researcher at Nokia Bell Labs, working on the intersection of privacy, machine learning and communications. He has been involved in several European flagship projects on wireless communications such as Hexa-X and mmMAGIC.

NUWANTHIKA RAJAPAKSHA is a Doctoral Researcher at the Centre for Wireless Communications, University of Oulu, Finland. She received the B.Sc. (Hons.) degree in electronics and telecommunication engineering from University of Moratuwa, Sri Lanka, in 2014, and the M.Sc. degree in wireless communications engineering from University of Oulu, in 2019. From October 2021 to June 2022, she completed a doctoral internship at Nokia Bell Labs, Stuttgart, Germany, where she worked on machine learning applications in physical layer. She is currently working under the 6G Flagship Programme in University of Oulu, and actively involved in the AI-driven air interface design task in Hexa-X EU Project and its extension Hexa-X II. Her research interests are machine learning for physical layer signal processing and resource allocation for cellular and cell-free massive MIMO.

ALESSANDRO RENDA received the M.Sc. degree in Biomedical Engineering from the University of Pisa, Italy, in 2017 and the Ph.D. degree in Smart Computing, jointly awarded by the Universities of Florence, Pisa and Siena, in 2021. Currently, he is an assistant professor at the Department of Information Engineering of the University of Pisa, Italy. His research interests include machine learning algorithms for data streams, federated learning and applications of deep learning methodologies.



MIKKO A. UUSITALO (Senior Member, IEEE) received the M.Sc. (Eng.) and Dr.Tech. degrees in 1993 and 1997, respectively, and the B.Sc. (economics) degree in 2003, all from predecessors of Aalto University. Since 2000, he has been at Nokia with various roles, including a Principal Researcher and the Head of International Cooperation at Nokia Research. He is currently the Head of the Research Department on Radio Systems Research Finland, Nokia Bell Labs, Finland. He is leading the European 6G Flagship Projects Hexa-X and Hexa-X-II. He has more than 80 granted patents or patent families and roughly same amount in the application phase. He is a Founding Member of the CELTIC EUREKA and WWRF, the latter one he chaired, from 2004 to 2006. He is a WWRF Fellow.