Editorial
Energy Efficiency of Machine-Learning-Based Designs for Future Wireless Systems and Networks

I. Introduction

WHILE 5G standards are being developed, research is moving toward designing the next generation of communications (e.g., 5.5G and 6G) which are expected to provide data rates of the order of 1 Tb/s using frequency bands in the range of 100 GHz to 3 THz. In addition to providing massive capacity and connectivity by exploiting new network architectures (e.g., cell-free massive MIMO, integrated terrestrial-aerial-underwater networks), radio transmission technologies (e.g., THz communications) and resource management techniques (e.g., end-to-end network resource slicing), future networks will support new context-aware applications and services (e.g., those based on joint communications and sensing) and provide connected intelligence in the era of Internet-of-Everything.

Recently, the usage and use cases of data-driven Artificial Intelligence (AI)/Machine Learning (ML) in communications systems have grown exponentially. The availability of large datasets and the reduction of costs of Graphics Processing Units (GPUs) are among the key factors for the development of AI/ML methods. Along with these, availability of flexible open source frameworks (e.g., Tensorflow, Pytorch, Caffee) has opened the doors for many applications of AI and ML in various areas [1]. In 6G, transformation of communications systems and networks will be from connected things to connected intelligence. The current trend in wireless communications systems design is to empower the communication devices with intelligence by exploiting their computational capabilities as well as storage hardware. This will enable implementation of complex algorithms to improve performance, reliability, and management of the network. However, this new trend imposes challenges due to heterogeneity of the devices (e.g., in terms of computational resources and battery power limitations).

In general, ML techniques for wireless systems have not yet reached the state of maturity. Some research works have shown that higher performance can be achieved with data-driven AI/ML-based methods when compared to classical methods; however, these methods often have deployment issues. In [2], the authors defined ten challenges in using ML techniques in 6G. These include learning efficiency, standardization, computation overhead, scalability for global intelligence, and tradeoff between learning accuracy and network overhead. The use of ML in 6G was discussed in the white paper [3] where the authors discussed the advances of ML in wireless communications in various layers and issues in the deployment of AI/ML-based methods. In [4], challenges in the use of ML models for resource management in IoT networks, such as the cost of training in ML and Deep Learning (DL) models and the lack of multitasking capabilities were addressed. The authors in [5] highlighted the computational costs of Deep Reinforcement Learning (DRL). They motivated the use of shallow machine learning and imitation learning in some scenarios where power and computation capabilities of the devices are limited.

In the rest of this article, we discuss different ML paradigms that have been used in wireless communications systems and we emphasize the need of new performance metrics related to energy-efficiency of AI/ML-based methods considering their time and space complexities. To this end, we review the usage of ML techniques in the PHYsical layer (PHY) as well as higher-layer and end-to-end design of wireless systems from energy-efficiency point of view, before we conclude the article.

II. Machine Learning Paradigms, Resource Constraints, and Energy Efficiency

A. ML Paradigms

Generally, ML techniques are categorized into three classes: (i) supervised learning, where an agent learns from data collected from real scenarios (or simulations of real scenarios), (ii) unsupervised learning, where the agent tries to find patterns from the data or description of the scenario, and (iii) Reinforcement Learning (RL), which is based on interacting with the environment and trial and error. These three paradigms have been used to solve different wireless communications problems. Also, ML techniques can be classified as either (i) a centralized ML technique, or (ii) a distributed ML technique. In wireless networks, Distributed ML (DML) techniques [6] have been widely used to solve well-known problems such as power control, spectrum management, user association, and computation offloading for edge cloud computing. Since the communication devices may generate a huge amount of data, distributed learning approaches are desirable to reduce the communication overhead and achieve efficient spectrum consumption.

We can distinguish two categories of DML [6], namely, *centrally-coordinated DML* and decentralized DML. Again,

Digital Object Identifier 10.1109/TGCN.2021.3099580

one approach within the centrally-coordinated DML is based on a *Parameter Server* (PS) *Framework*. Basically, there are two groups of network entities: servers and clients. A server allocates data to the clients, maintains the global model parameters, and aggregates the updates for the global model parameters obtained from the clients [7]. Another approach within the centrally-coordinated DML category is *Federated Learning* (FL) [8], where data are generated locally at the clients. This approach reduces the amount of transmitted data, because the server does not need to send data to the clients, and also only the model updates are transmitted to the server by the clients. Training the machine learning models in a distributed manner using the FL approach is attractive, because data are generated locally on the devices, and for communication and privacy reasons it may be infeasible to transfer the data to the cloud for training. Therefore, the training is distributed on the devices and performed locally. Also, FL is more attractive from the perspective of energy efficiency compared to the centralized solution when the data size is much larger than the number of parameters in the model.

Both of the PS and FL approaches can be referred to as a *data parallel* approach. As an example, MapReduce, which is a software framework for distributed processing of large data sets on computing clusters, was adopted in wireless to reduce the computational requirements of the learning process. Motivated by the new generation of devices with an increase of storage and computational capacities, MapReduce was used in mobile edge computing [9], [10]. Another approach within the centrally-coordinated DML category is *Partitioned Learning* that works in a setting where a supervised model is divided into several blocks of different parameters and each block is downloaded to different devices for learning. The outcome of the learning, i.e., the model parameters are returned to the server to refresh the global model. This is referred to as a *model parallel* approach.

The second category of DML is *centerless DML* where devices/agents learn separately to achieve a global goal. One of the promising methods in this category is Multi-Agent Reinforcement Learning (MARL) which is an extension of single-agent reinforcement learning to a multiple agent setting. The MARL problems are formulated as stochastic games and the agents learn collaboratively or competitively to jointly solve a specified problem or achieve the optimal global solution.

B. Resource Constraints in AI/ML-Based Design

To enable AI/ML capabilities on small devices, the ML algorithms need to provide high model accuracy while working within the resources constraints of the devices [11]. The major resource constraints to run ML algorithms in the devices are described below.

1) Processing Speed: The processing speed in a computing hardware determines the system response in terms of *throughput* (i.e., rate at which input data is processed) and *latency* (i.e., time taken to process a single input to its output). It is commonly measured in clock frequency (i.e., the number of cycles per second) of a processor. For ML algorithms, there are two main metrics for runtime of AI-based methods: FLoating Point Operations Per Second (FLOPS) and Multiplier-Accumulate (MAC). MAC is the number of multiplications and additions done in the algorithms. In some architectures, the processing speed of the system is affected by communication latency (e.g., in a distributed computing paradigm).

2) Memory: Memory is one of the major resources required by the ML algorithms because it stores the data to be used for learning. Moreover, storing the model parameters and variables takes a significant amount of memory during the update. The memory footprint of an ML method is determined by the required storage capacity and the speed of data access. Besides the model and the data size, querying the model parameters is also time and energy consuming because a single MAC operation takes three memory reads and one memory write.

3) Power Consumption: Power consumption is a critical factor for on-device learning. The energy consumed for running an algorithm depends on various factors, e.g., runtime, memory access and computations on data. In general, the required energy for computation is linearly proportional to the size of the data set. Due to the non-deterministic nature of power consumption of AI/ML models, to evaluate a model, an estimate of power/energy usage is performed through a surrogate function that depends basically on the runtime and memory.

C. Energy-Efficiency as a Key Performance Metric

ML has been used as an alternative to classical methods to solve "wireless" problems such as those related to spectrum management, beamforming, user association, multiple access and resource allocation. As inherited from other fields, the main performance metric used is the accuracy of the model. However, due to the specific nature of the wireless communications systems and the characteristics of the wireless communications devices, accuracy is not a sufficient metric. For an ML model, since the training and inference phases can consume high computational resources and large storage capacity for the data, other metrics such as computation overhead and required storage should be studied as well. In particular, IoT devices are limited in power and computational resources. To compare the ML-driven approaches and their feasibility of deployment in these devices, an accurate estimation of the real-time energy consumption would be required.

Modeling energy consumption for machine learning algorithms at software and hardware levels was reviewed in [12]. At the software level, the energy consumption models use the characteristics of the algorithmic properties (e.g., kernel size of neural network) and the types of the instructions contributing to different levels of energy consumption. At the hardware level, the models consider components such as processor, memory, and I/O peripherals that contribute to energy consumption during training and execution modes. Empirical measurements show that reducing the number of weights in a deep neural network model does not necessarily lead to energy saving, and therefore, the number of weights is not a

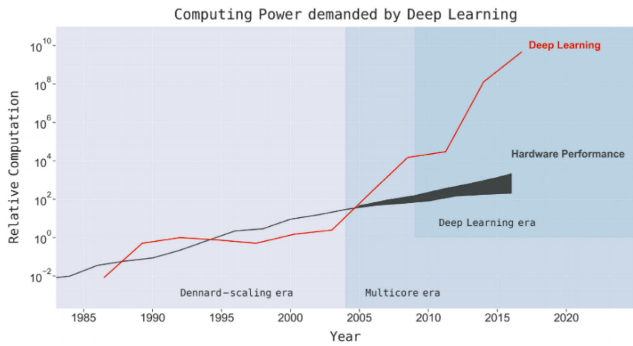

Fig. 1. Computer power used in deep learning models of all types [15].

good indicator of energy-efficiency of the model. For convolutional neural networks, a big part of energy is consumed by the convolution layer [13]. Using deep neural network energy estimation tool (https://energyestimation.mit.edu/), [14] proposed an approach to reduce energy consumption of convolutional neural networks.

As shown in Fig. 1, computing power demanded by DL techniques is exploding compared to the hardware improvements. The high computational requirements of DL and lack of accurate metrics of energy consumption and storage requirements are challenges that need to be overcome for successful and widespread deployment of DL models in wireless communications systems. Therefore, one important future direction of 6G research will be to develop, model, and analyze data-driven ML methods considering computational efficiency and energy efficiency.

III. Energy-Efficiency in Designing ML-Based 6G Systems

A. *Energy-Efficiency in ML-Based PHY Design*

ML has been used in physical layer (PHY) design of wireless communications systems due to limitations of the mathematical modeling of complex environments and/or due to algorithmic complexity of solving optimization problems for real-time implementation. For instance, DL was applied for channel coding [16], [17] to enable rapid coding and decoding for low-latency services. Also, DL can be employed to overcome the limitations of traditional techniques in presence of non-line-of-sight propagation.

Many researches took advantage of ML techniques in channel estimation process for wireless networks. The Linear Minimum Mean Square Error (LMMSE) method provides an optimal channel estimation for linear and stationary channels; however, real channels could be non-linear and non-stationary. Therefore, DL is a promising approach to solve this issue since it can be trained on complex channels [18]. The neural network should be trained offline because of the requirement of long training on different samples of different channel models, and therefore, it requires high computational resources.

In [19], a DL-aided non-orthogonal multiple access (NOMA) system was introduced to replace the traditional methods that have high computational complexity. In the physical layer, ML impacts the design of hardware in different levels [3]. It was proposed for substitution of specialized functions (e.g., channel estimation) and updating of modules that are not currently well-solved due to some nonlinearities. Another use of ML techniques is the joint optimization of different modules in the physical layer. For instance, signal decoding and waveform design can be optimized together. However, the complexity at the receiver would be too high; therefore, a sub-optimal end-to-end mapping mode can be learned by a neural network. Besides, methods that combine ML techniques and model-based methods are promising to overcome the defects of ML-based methods such as requirement of long training data, slow convergence, underfitting, and overfitting. Thus, ML has opened the doors for designing wireless PHY in a different way.

Generally, accuracy of the model is used as a metric to judge the performance of ML-based techniques. However, since hardware constraints will play a critical role in 6G [20], the computational overhead and storage requirements should be considered as well in the design of ML-based techniques. Also, high power consumption of signal transmission/reception in the radio transceivers (e.g., those operating in millimeter-wave and Terahertz bands) will affect PHY design and algorithms. Besides, for deep learning in the physical layer, online updates are necessary to handle factors that were not considered in the offline training. The performance of AI/ML-based methods needs to take into account the energy consumption during training and inference phases.

The process of deploying AI/ML-based methods in embedded wireless devices goes through software simulations, prototyping, and production phase. The training process is done in the prototyping phase on specialized FPGAs for neural networks that allow high flexibility and speed in the design. Since the hardware is different, there is a gap between simulation phase/prototyping and production phases [3]. Most of the ML-based algorithms designed for PHY layer are still in the simulation phase. Nonetheless, an early study in the simulation phase on the power and storage consumption is essential to develop energy-efficient methods for PHY technologies.

Reducing the model complexity is one of the resource-efficient approaches that can be considered while building the model to adapt the model architecture to the available resource constraints in the target device. It consists of adding constraints on the parameters of the model to control the memory footprint and the computational complexity of the ML algorithm [11].

B. *Energy-Efficiency in ML-Based Radio Link/Network Layer Design*

ML techniques can be used in multiple access/radio link layer design problems such as user selection, power management and resource allocation due to the high computational complexity of the classical optimization-based methods. Many radio link/network layer problems are combinatorial and/or non-convex optimization problems; therefore, heuristic algorithms are used to find sub-optimal or near-optimal solutions. ML techniques are useful to obtain low-complexity solutions that can be implemented online. DRL is suitable for decision making and control problems in large-scale and complex networks. For instance, in [21], a DRL-based medium access control protocol was proposed for a heterogeneous wireless network. DRL was used in [22] to address the problem of beamforming design for multiple access in cell-free MIMO networks. DRL was also used in [23] to design a MIMO-based full-duplex energy harvesting system. In [24] a Deep Q-Learning (DQL)-based approach was proposed for power allocation. The performance metrics however consider only the accuracy of the model and do not consider the energy consumption. Since energy consumption for the learning process of the neural networks may be high for IoT devices, where the battery life is limited, the trend now is to design ML-based methods that perform both offline and online learning. Also, ML methods need to consider the hardware capabilities of the devices, because some devices may not be able to perform learning due to the limited computational resources (e.g., smart watches). An approach that can be used to reduce the memory usage on the edge-devices and enable fast inference is model compression [25]. Two broad approaches for doing this are: i) *quantization*, which reduces the precision of the parameters values and the memory footprint, hence it makes the computations faster, and ii) *model pruning* which decreases the number of the model parameters to improve the storage and computational time. SqueezeNet [26] is an efficient neural network used for resource-constrained devices that uses deep compression by weight pruning, quantization, and Huffman encoding. But the decrease of storage and computation comes at the cost of accuracy. Network layer design in a software-designed networking environment can also exploit DML techniques (e.g., Federated Learning). Since a distributed training can suffer from communication overhead, one major research challenge is to reduce the communication rounds needed for convergence to its lowest possible.

C. Energy-Efficiency in ML-Based End-to-End Design, Network Management, and Services

AI/ML techniques can be used for end-to-end design, e.g., for network resource slicing and virtualization on an end-to-end basis. Also, ML can be conveniently used (e.g., in a Software-Defined Networking [SDN] environment) for anomaly detection, prediction of performance trend and Key Performance Indicator (KPI) degradation in a network. The goal of using ML in such scenarios is to keep the KPIs within predefined thresholds. Using machine learning based on the data collected from user equipments and base stations in the network will enable automated network management and control. Furthermore, ML techniques can be used to solve the tasks of traffic prediction and traffic classification. In [27], a guideline for the application of ML in networking was provided for traffic prediction and classification. ML techniques have been used in the management of unmanned aerial vehicles (UAVs) in a wireless communication environment. In a vehicular communications environment, ML approaches, when used for the selection of network interfaces based on the anticipated resource efficiency, should consider energy efficiency of the network. At the same time, ML-based solutions themselves should be energy-efficient, and the related implementation challenges should be considered in the design of these solutions.

Machine learning has many applications in a variety of domains such as robotics, computer visions, natural language processing, and healthcare. These applications will be supported in future 6G wireless systems and networks. ML techniques will be used to ensure the accessibility and availability of the corresponding services and meet their response time requirements. Due to the need for high-speed processing of massive data and the stringent response delay requirement, there is a gap between the current hardware of the devices and the AI/ML-driven computational requirements. The use of parallel, multicore, multithreaded, GPU and Tensor Processing Unit (TPU) chips will become the foundation of future generation computing and communications devices and systems [28].

IV. CONCLUSION

The existing research works on applications of ML techniques in 6G networks lack rigorous investigations on energy consumption of ML methods in practical scenarios. For large models, the ML algorithms are computation intensive while the wireless devices are generally limited in power, storage, and computational resources. Although accuracy of the AI/ML models is an important metric, memory footprint, computational complexity, and power consumption are other metrics that need to be considered in the design of practical ML methods. Lightweight and compressed models that are known for their energy-efficiency would be preferable for wireless devices. Also, profiling tools should be used while designing ML-based approaches on real scenarios of large networks to define the computation bottlenecks and reduce the energy consumption. Designing energy-efficient AI/ML-based methods for 6G communications systems and networks is a fertile area of research.

EKRAM HOSSAIN
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, MB R3T 2N2, Canada

FIRAS FREDJ
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, MB R3T 2N2, Canada

REFERENCES

[1] Y. Bengio, Y. Lecun, and G. Hinton, “Deep learning for AI,” *Commun. ACM*, vol. 64, no. 7, pp. 58–65, Jul. 2021. [Online]. Available: https://doi.org/10.1145/3448250

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

[3] S. Ali *et al.*, “6G white paper on machine learning in wireless communication networks,” 2020. [Online]. Available: https://arxiv.org/abs/2004.13875.

[4] 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 Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2nd Quart., 2020.

[5] F. Tang, B. Mao, Y. Kawamoto, and N. Kato, "Survey on machine learning for intelligent end-to-end communication towards 6G: From network access, routing to traffic control and streaming adaption," *IEEE Commun. Surveys Tuts.*, early access, Apr. 13, 2021, doi: 10.1109/COMST.2021.3073009.

[6] S. Hu, X. Chen, W. Ni, E. Hossain, and X. Wang, "Distributed machine learning for wireless communication networks: Techniques, architectures, and applications," 2020. [Online]. Available: https://arxiv.org/abs/2012.01489.

[7] Z. Zhang, L. Yin, Y. Peng, and D. Li, "A quick survey on large scale distributed deep learning systems," in *Proc. IEEE 24th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Singapore, 2018, pp. 1052–1056.

[8] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for Internet of Things: Recent advances, taxonomy, and open challenges," *IEEE Commun. Surveys Tuts.*, early access, Jun. 18, 2021, doi: 10.1109/COMST.2021.3090430.

[9] F. Li, J. Chen, and Z. Wang, "Wireless mapreduce distributed computing," 2018. [Online]. Available: http://arxiv.org/abs/1802.00894.

[10] M. Ji and R.-R. Chen, "Fundamental limits of wireless distributed computing networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, 2018, pp. 2600–2608.

[11] S. Dhar, J. Guo, J. Liu, S. Tripathi, U. Kurup, and M. Shah, "On-device machine learning: An algorithms and learning theory perspective," 2019. [Online]. Available: http://arxiv.org/abs/1911.00623.

[12] E. García-Martín, C. F. Rodrigues, G. Riley, and H. Grahn, "Estimation of energy consumption in machine learning," *J. Parallel Distrib. Comput.*, vol. 134, pp. 75–88, Dec. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0743731518308773

[13] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Seoul, South Korea, 2016, pp. 367–379.

[14] T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 5687–5695.

[15] N. C. Thompson, K. H. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2020. [Online]. Available: https://arxiv.org/abs/2007.05558.

[16] E. Nachmani, Y. Beery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *Proc. 54th Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, Monticello, IL, USA, 2016, pp. 341–346.

[17] A. Askri and G. R.-B. Othman, "DNN assisted sphere decoder," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, 2019, pp. 1172–1176.

[18] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[19] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[20] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[21] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1277–1290, Jun. 2019.

[22] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1028–1042, Apr. 2021.

[23] Y. Al-Eryani, M. Akrout, and E. Hossain, "Antenna clustering for simultaneous wireless information and power transfer in a MIMO full-duplex system: A deep reinforcement learning-based design," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2331–2345, Apr. 2021.

[24] K. I. Ahmed and E. Hossain, "A deep Q-learning method for downlink power allocation in multi-cell networks," 2019. [Online]. Available: http://arxiv.org/abs/1904.13032.

[25] M. G. S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, and F. Hussain, "Machine learning at the network edge: A survey," 2019. [Online]. Available: http://arxiv.org/abs/1908.00080.

[26] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2016. [Online]. Available: https://arxiv.org/abs/1510.00149.

[27] M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, "Machine learning for networking: Workflow, advances and opportunities," *IEEE Netw.*, vol. 32, no. 2, pp. 92–99, Mar./Apr. 2018.

[28] H. Yao, T. Mai, C. Jiang, L. Kuang, and S. Guo, "AI routers & network mind: A hybrid machine learning paradigm for packet routing," *IEEE Comput. Intell. Mag.*, vol. 14, no. 4, pp. 21–30, Nov. 2019.

Ekram Hossain (Fellow, IEEE) is a Professor and an Associate Head (Graduate Studies) with the Department of Electrical and Computer Engineering, University of Manitoba, Canada (http://home.cc.umanitoba.ca/~hossaina). His current research interests include the design, analysis, and optimization of wireless networks with emphasis on beyond 5G cellular networks. He received the 2017 IEEE ComSoc Technical Committee on Green Communications & Computing Distinguished Technical Achievement Recognition Award "for outstanding technical leadership and achievement in green wireless communications and networking." He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2017–2020. He currently serves as the Editor-in-Chief for IEEE PRESS. He served as the Editor-in-Chief for IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2012 to 2016. He was elevated to an IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks." He is a member (Class of 2016) of the College of the Royal Society of Canada and a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada.

Firas Fredj received the B.Sc. degree in engineering from the Ecole Polytechnique de Tunisie in 2020. He is currently pursuing the M.Sc. degree in electrical engineering with the University of Manitoba, Canada. His research interests include the design of wireless communications systems and networks using machine learning techniques.