COGNITIVE CUSTOMER PROFILING AT THE EDGE: CHALLENGES AND RESEARCH DIRECTIONS

Karima Velasquez[1], David Perez Abreu[2,1], Pedro Neves[3], Miguel Santos[3], Fernando Santiago[3], Nuno Lourenço[1], Marilia Curado[1], Edmundo Monteiro[1], Carlos Amaro[3]
[1]University of Coimbra, CISUC, Department of Informatics Engineering, Coimbra, Portugal, [2]Instituto Pedro Nunes, Laboratory for Informatics and Systems, Coimbra, Portugal, [3]Altice Labs, Aveiro, Portugal

Corresponding author: Karima Velasquez, kcastro@dei.uc.pt

Abstract – *The proliferation of automated services and applications has generated large amounts of data that must be handled with artificial intelligence and machine learning techniques, impacting businesses in different fields. Cognitive customer profiling services allow a better user experience enabling the tailoring of the customer experience to specific consumer patterns and preferences. Cognitive customer profiling services require heavy processing power in order to get accurate results from the artificial intelligence/machine learning models, which has led to services taking advantage of the cloud-to-edge continuum. This trending approach brings a set of challenges that have to be addressed. The more recent edge intelligence paradigm introduces artificial intelligence capacities to edge nodes, enabling a new set of solutions. However, artificial intelligence techniques must be adapted to deal with the specific characteristics of these scenarios. This paper identifies the challenges of running cognitive customer profiling in centralized scenarios, such as the cloud, to later introduce other approaches more suited for scenarios like the edge, identifying challenges and future research directions in the area.*

Keywords – artificial intelligence, cloud-to-edge continuum, cognitive services, edge intelligence, user profile.

1. INTRODUCTION

As more significant amounts of data are generated due to the increase of automated services and applications, Artificial Intelligence (AI) and Machine Learning (ML) have become increasingly crucial. Cognitive services bring AI within reach of different aspects of our daily lives, impacting how we work, do business, entertain ourselves, and interact with the world [1]. Different application domains can leverage AI techniques to provide high-performance cognitive services, such as speech recognition, image classification, healthcare, and recommendation systems. AI has also been used for profiling in different fields including health [2], food [3], and customer information [4] [5], allowing a personalized service tailored to customers' needs and preferences.

The disruptive growth of data represents a challenge for cognitive services since the profiling of an entity (e.g., device, service, or user) requires a distributed and multipoint engagement to consider data granularity, media diversity, and service-side solutions [6]. To cope with these issues, it is necessary to consider how to aggregate and integrate data from different sources, as well as possible solutions to mitigate a potential risk when handling individual privacy information.

Cloud environments arise as a centralized solution that enables access to a set of computational resources on demand, from any device, and at any time, accelerating industry dynamics and fueling digital transformation [7]. To deal with these copious amounts of data, AI models become more complex, which leads to a need for increasing the computational resources available for their training and execution. Thus, the cloud becomes an appealing alternative that offers the necessary resources on demand and in a flexible manner. Nonetheless, low latency has become a critical aspect for recent applications such as assisted driving and augmented/virtual reality. Relying on the cloud for the execution and training of the models increases the response time for delay-critical applications.

Furthermore, the exchange of sensitive data between its original source, usually at the periphery of the network until its core, where the cloud resides, incurs potential privacy risks and increases in latency and response time. Therefore, other paradigms emerge, such as the edge, a distributed scenario that offers computational resources closer to the user, appearing as an extension of the services offered by the cloud [8]. Nevertheless, the characteristics of the edge differ from those in the cloud, being more resource-constrained and also of a distributed nature. This poses new challenges that have to be addressed before considering the edge as a suitable location for the execution of AI models, speeding up the processing time and reducing privacy risks.

The cloud-to-edge continuum paradigm envisions an orchestrated use of resources and services along each of the tiers (i.e., cloud, fog, and edge), allowing the evolution of the communication infrastructure and its applications to enhance the Quality of Experience (QoE) of the users [9]. The vast proliferation of devices in the cloud-to-edge and the scattering of data sources lead to the exploration of new data analytics approaches, which aim to achieve proficient knowledge of the environment in order to make smart decisions.

This paper explores the challenges of executing AI-based solutions for customer profiling in the cloud, while providing alternative research paths, and suggesting tailored AI techniques for more specific execution domains, such as the edge. Additionally, an analysis and complete discussion of the cloud-to-edge continuum in this matter is presented, as well as its benefits and implications from the Cognitive Customer Profiling (CCP) point of view. The contributions of this paper are the following:

- A taxonomy of AI in the cloud-to-edge continuum (see Section 2);
- A list of challenges for centralized CCP services (see Section 3);
- A set of recommendations for AI techniques that can be applied at the edge of the network to support CCP services, based on Edge Intelligence (EI) (see Section 4).

This paper is organized as follows. Section 2 presents the background needed, including a description of a novel approach that combines the edge computing paradigm with AI, called *Edge intelligence*, and a brief definition of profiling via AI. Section 3 describes the particular use case of CCP and the challenges it presents. Section 4 highlights how edge intelligence can help overcome profiling use case issues. Finally, Section 5 wraps up the paper and presents future research directions.

2. BACKGROUND

This section describes the concepts of edge intelligence and user profiling via AI, which will be necessary for the case study introduced in the following section.

2.1 Edge intelligence

With Cisco estimating 3.6 networked devices per capita by 2023 [10], and with the increasing requirements from autonomous and smart applications that become more popular each day, the edge becomes an appealing scenario to support the proliferating use of AI-based solutions for intelligent applications, paving the way to EI; a booming paradigm that integrates edge computing and AI. EI refers to the capacity to analyze data locally at its collection location, instead of relying on the cloud for it [11]. EI includes intelligent offloading to edge servers, collaboration between edge servers via federated mechanisms, and running ML models at the edge nodes to perform analysis services and user behavior, among others; all without the need to use cloud resources.

Typical use cases for EI include [12]:

- Connected vehicles: allowing real-time monitoring and collective perception;
- Smart factories: enabling monitoring and control of assembly lines;
- Real-time gaming: facilitating virtual reality gaming at lower latency, with support of low-cost end-user devices;
- Diagnostics for healthcare: allowing the processing of sensitive data at the proximity of end users and providing real-time diagnostics; and
- Profiling: analyzing service and/or user data at the edge of the network infrastructure, improving data privacy and response time.

Two aspects in which EI can be applied. The first aspect refers to the zero-touch management functions for the network infrastructure: *Intelligence for the edge*. AI will be necessary to perform orchestration tasks in such a dense scenario where human interaction is not feasible. AI-based solutions will enable a better analysis of vast volumes of data, extracting insights that empower smart decision-making for orchestration and management purposes.
The second aspect to which EI can be applied is related to the training and execution of models within the edge frontier: *Intelligence on the edge*. This refers to a framework for running training and inference processes of AI models, extracting information from massive and distributed data available at the edge. Naturally, it is possible to combine both approaches: *Intelligence for the edge at the edge*, where the edge nodes execute ML models aimed at edge orchestration functions.
The ML lifecycle involves two phases: *training and inference*. During the training phase, the model is fed with a selected dataset to enable it to acquire a comprehensive understanding of the data it will later analyze. Subsequently, in the inference phase, the model is expected to generate accurate predictions using real time data [13]. According to the location within the network infrastructure where the execution of the ML models takes place, it is possible to create a taxonomy of AI in the cloud-to-edge continuum, depicted in Fig. 1:

1. *Cloud intelligence*, where both the training of the models and final inference take place in the cloud, in a centralized manner;
2. *Cloud training + Continuum inference*, in which it is possible to take advantage of the vast cloud resources for heavy training tasks, while the inference process occurs in the cloud-to-edge continuum;
3. *Continuum intelligence*, where all the involved processes (i.e., training and inference) are performed inside the continuum;
4. *Continuum training + Edge inference*, in which the training can take place anywhere within the continuum, but the inference is executed solely on the edge; and
5. *Edge intelligence*; where both training and inference take place on the edge devices.

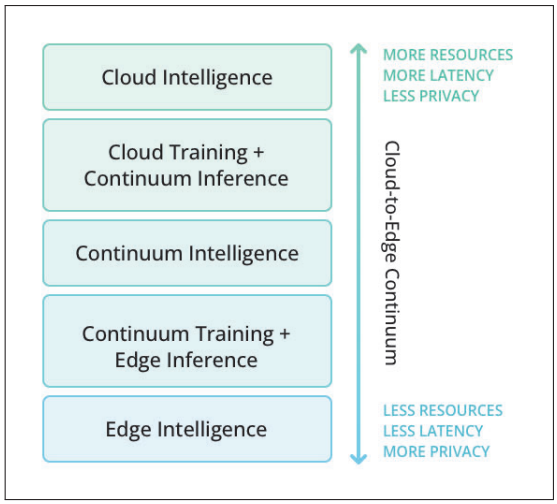


Figure 1 – AI in the cloud-to-edge continuum taxonomy

In the case of tasks that occur in the continuum or directly in the edge, it is also possible to use federative approaches, by having different nodes collaborate, taking advantage of the distributed nature of these infrastructures.

2.2 Cognitive customer profiling

Data analytics empowered by AI are revolutionizing the way businesses interact with their customers, improving the customer experience by leveraging the knowledge about their preferences and consumer patterns [5]. Personalization and enhanced engagement based on contextual and behavioral data have been improved due to the use of AI in the way of AI chatbots, content generation, and customer insights.
AI has been used to create customer profiles during their visits to stores including gender, age, personality, emotions, and products they interacted with, allowing a personalized experience for future visits as well as improving recommendation systems [4]. Learning from previous behaviors makes it possible to predict future preferences [14]. This profiling of customers based on AI is referred to as *CCP*.
Some of the AI techniques used in CCP are fuzzy techniques [15][16], evolutionary algorithms, transfer learning, neural networks and deep learning [4], and active learning. To deal with the significant amount of data needed to achieve accurate profiles, these solutions traditionally rely on cloud data centers and their processing power.
The following section describes a particular case of the CCP services, related to digital service providers.

3. COGNITIVE CUSTOMER PROFILING AT THE EDGE

This section describes the usage of CCP services, an initial centralized vision, and its challenges, and how the edge can provide a distributed approach to overcome them.

3.1 CCP services

One of the key areas of profiling customers is knowledge. Customer knowledge represents an important added value as a support for business decisions. For example, in the customer care area, it allows adapting customer communications based on their preferences and selecting the best channel/moment to establish contact. In the TV domain, the knowledge of the customer allows the recommendation of the most appropriate content. Another example is in the digital marketing area, allowing the marketing of services in a more personalized way, thus promoting up/cross-sell.

To explore these business scenarios, it is paramount to build a 360° customer view, including both direct and inferred customer profile dimensions. The direct component of the profile is obtained directly through other existing platforms in the service providers' ecosystem. The inferred component of the customer profile is supported by cognitive mechanisms, such as AI and Data Science (DS) techniques, to infer a set of customer-related insights. Some examples are: preferred viewing TV content (e.g., science fiction content, cooking, sports), preferred leisure activities (e.g., theater, movies, sports), customer sophistication level (e.g., early adopter), preferred contact time and channel (e.g., night via SMS, morning via WhatsApp), sentiment analysis (e.g., irritable, patient). The complexity of the AI/DS algorithms, such as decision trees and neural networks, depends on the type of data available and the nature of the use case we intend to undertake.

To address this need, CCP services are being designed and developed. Briefly, the CCP services provide an advanced customer profile view, thus allowing the exploration of the business scenarios presented above. To this end, exposing the CCP services to other platforms is essential, either through direct requests or by sending notifications based on previously defined thresholds. For example, marketing automation platforms will consume the CCP services to incentivize customers with fully personalized marketing campaigns. Another example is customer care platforms, which will be able to use the advanced customer knowledge provided by CCP services to improve the customer experience in call centers.

In the next section, we discuss the main challenges that must be considered in the case of thinking about performing a centralized deployment or instantiation of CCP services.

3.2 CCP services centralization challenges

The centralized deployment of the CCP services platform in a digital service provider's architecture raises critical challenges that may jeopardize its use and, above all, the expected impact on the business value chain. The following paragraphs depict the **three main challenges** that a centralized deployment of this platform poses.

The **first challenge**, and certainly one of the most relevant, is the **impact on the communication network of the digital service provider**. The CCP services platform core is based on AI profiling models, which require a large amount of input data to be executed and produce the expected profiling insight. Additionally, it is expected that the AI-based profiling models will have to be executed with high frequency to update and optimize the customer profile dimensions that are being inferred. Therefore, huge amounts of data will have to circulate from the customer devices towards the CCP services platform located in the core network, putting a lot of pressure on the operator's network and, ultimately, even compromising other services that are being provided.

The **second challenge** identified is related to the **real time generation of customer profiling insights**. Transporting the required input data towards the AI-based profiling models running on the CCP services platform, located in the core network, is very time-consuming. Therefore, in scenarios that require real time predictions and/or recommendations generation, this delay can compromise the use of the insight to impact the business. An example is TV content viewing recommendations. In this case, a personalized viewing recommendation must be delivered in real time whenever a customer finishes viewing content. This means that a viewing recommendation request must reach the CCP services platform, and subsequently, the AI-based

model is executed and produces the recommendation. This must be achieved in real time, which is difficult to reach with a centralized platform.

The **third challenge** identified is related to **privacy violation**. The CCP services platform contains sensitive customer data that must not be accessed without a valid justification. Centralizing all the customer profiling information in a single location can be risky. For example, in case of a successful cyberattack on the platform, the probability of compromising the privacy of all the data that is centralized is high.

The main challenges of CCP services centralization were introduced and discussed in this section; below, we introduce and analyze possible ways to overcome them.

3.3 Edge-distributed CCP services

One solution to surmount the issues presented above is to distribute CCP services across the edge of the operator's network, as illustrated in Fig. 2. The figure shows the operator network segmented into three major areas. The first area is the access network through which end users obtain connectivity (i.e., fixed and/or mobile). Next is the backhaul network (i.e., fixed and/or mobile) in which data is transported from the customer access network to the core. Finally, the figure also depicts the core network in which the service platforms, network management platforms, and network control platforms are located.

As depicted in Fig. 2, CCP services are no longer centralized in the operator's network core, but distributed throughout the edge of the network. The figure, as an example, shows one CCP instance deployed in the core network, several CCP instances deployed in the backhaul segment (i.e., fixed and mobile), as well as several CCP instances located in the access network (i.e., fixed and mobile), very close to the end customer. Distributing the CCP services across the network edge will allow the profiling services to be consumed from these instances without having to reach the instance centrally deployed at the core network. Thus, transporting the data to the CCP instance located in the core network will not be necessary; consequently, the traffic flowing through the network will decrease considerably.

Reducing the network pressure as described previously, deploying the CCP services platform in the edge enables real-time scenarios. For example, taking into consideration the proximity of the CCP services and the profiling consuming information platform (e.g., TV content platforms), it is expected that the request/response interactions with the CCP services platform are faster and therefore encourage the exploitation of real-time business use cases. Finally, distributing the customer profile information across the various network nodes will reduce the cyberattack impact on sensitive profiling data. That is, if the data of one CCP instance located at the edge is compromised, the profiling data of the other instances will remain safeguarded.

Although the distribution of CCP services addresses the centralization challenges presented in this document, some critical challenges in the distributed approach must be considered. Namely, it is necessary to define the appropriate location for the various CCP services' instances that are distributed across the edge and, not less importantly, also define the appropriate sizing for these platforms. In other words, we do not want to place CCP instances at every location in the edge or over/undersize these instances. This would bring considerable costs in terms of hardware, as well as in the operation and maintenance of additional platforms.

Intelligent decision engines based on AI models require, on the one hand, to predict the requests that will be made to the profiling platforms and, on the other hand, recommend the edge locations where CCP service platforms should be deployed. Complementarily, mechanisms based on AI models are also being studied to predict which profiling information should be available in each CCP instance, thus allowing the sizing recommendation of these instances. Since the positioning of the CCP service instances and the profiling information required in each of them varies over time, intelligent orchestration mechanisms are also being studied to guarantee the coherence of the entire distributed profiling architecture, CCP service platform positioning, information, and sizing. Finally, an interlinking mechanism must also be designed to obtain profiling information that is not present in a given CCP service instance but only in a neighboring instance.

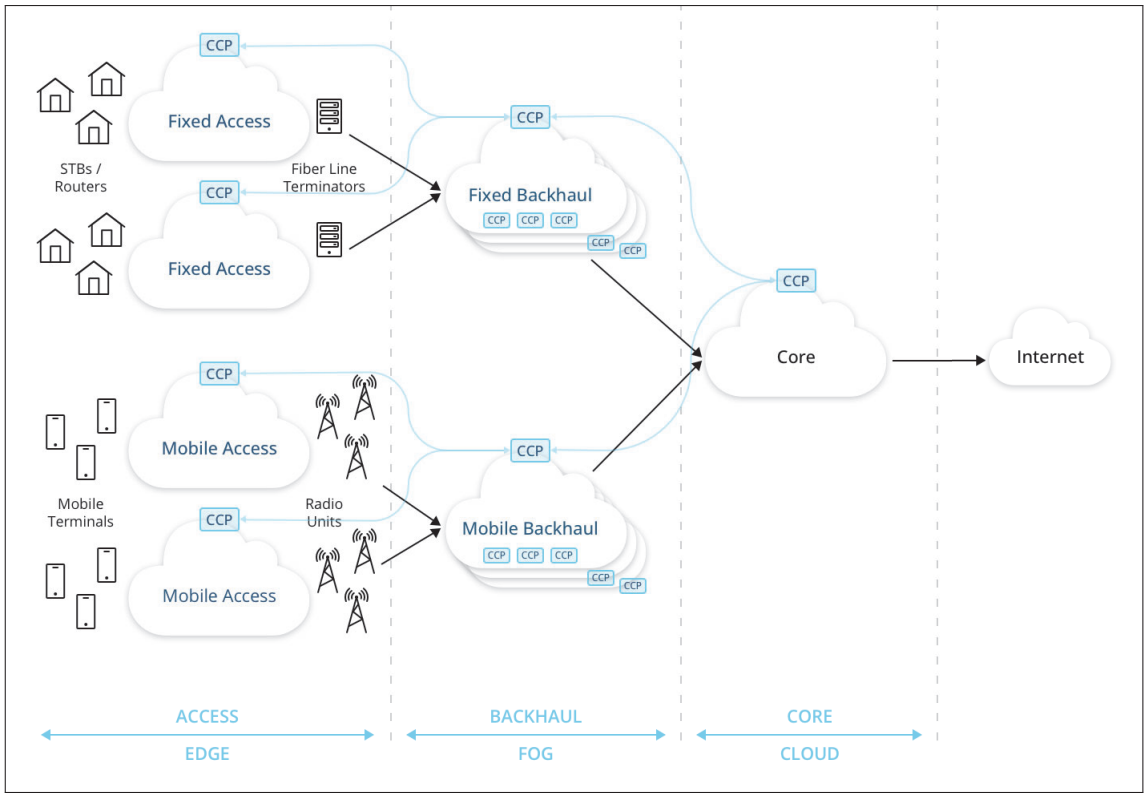


Figure 2 – Edge-distributed CCP services

4. HOW EDGE INTELLIGENCE MEETS CCP REQUIREMENTS

In the scenario described in Section 3, the highest volumes of data are concentrated in the end-user devices, which are resource-constrained by nature. Thus, a possibility is offloading pr eprocessed data to edge nodes that can perform some tasks that cannot be executed in the end devices, while more demanding tasks will, in turn, be offloaded to the cloud but receive only a part of the data available at the edge nodes [12].

The CCP services centralization challenges identified in Subsection 3.2 can be overcome with *EI*. The first challenge was related to communication impact. By keeping the data collection and model execution at the edge, without reaching the network core in most cases, it is possible to reduce the network traffic for service providers. The second challenge was associated with the real-time execution of the profiles. The propagation delay would be significantly reduced by running the models at the edge; furthermore, it has been proven that executing services at the edge instead of the cloud directly impacts the latency [17] [18]. Finally, the third challenge was related to data privacy. The customer profile is built from sensitive data; thus, it becomes crucial to keep this data closer to the user, at the edge level, avoiding unnecessary travel across the network where more security breaches could happen.

At the perimeter of the cloud-to-edge continuum, the classical approach of ML is traditionally not suitable taking into consideration that despite the availability of a high volume of data in the lower tiers, the capability of processing data is limited given the source limitations of the devices at this level [12]. To deal with this characteristic, preprocessed data can be offloaded to edge devices to perform tasks that cannot be executed by Internet of Things (IoT) devices placed in the edge. This process can be done until reaching the cloud tier, where more complex models can be executed but to keep better data privacy and lower network congestion, only a part of the data from the IoT is available. Two challenges arise to evolve AI/ML for the cloud-to-edge continuum: defining a trade-off between accuracy and resource demand, and utilizing Federated Learning (FL) of the edge and IoT devices deployed in the edge.

Regarding the trade-off between accuracy and resource demand, it is important to highlight the lim-

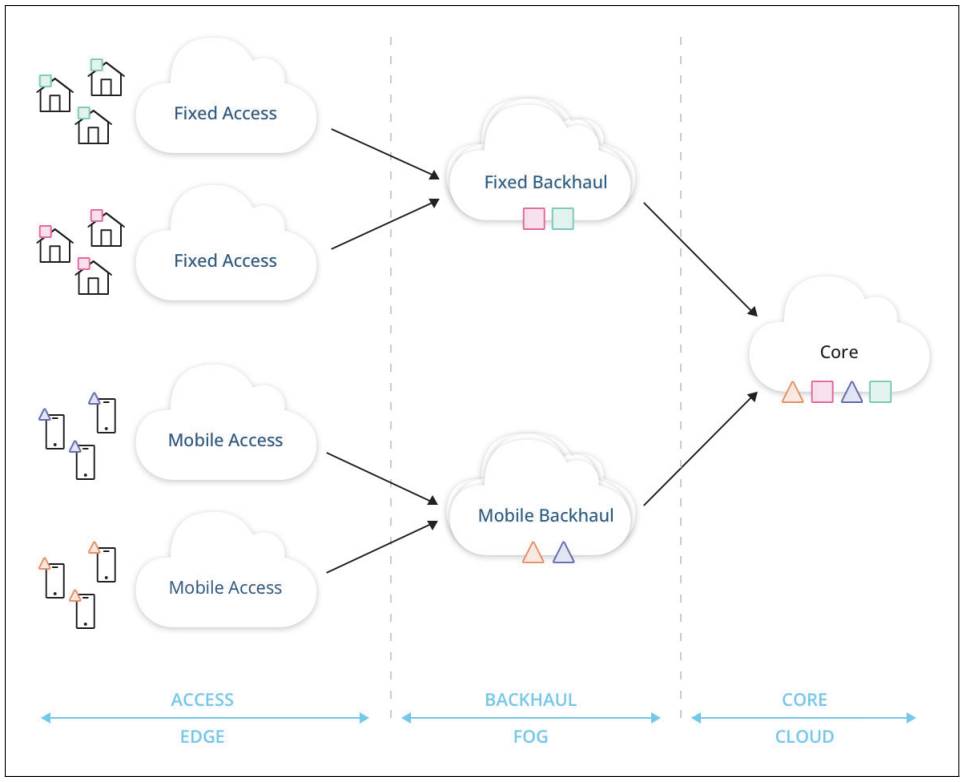


Figure 3 – Federated learning

itations of IoT and edge devices in comparison with those in the cloud, making these latter ones more suitable for more complex ML models. Although there is a common goal to achieve the highest possible accuracy, other metrics to take into consideration are related to resource usage (e.g., energy, memory) and service performance (e.g., latency). A research venue is exploring the elasticity of the models (i.e., adapting their size) to be able to adapt to available resources. Some techniques that can be used in this context are the miniaturization of models using model pruning, trans-precision, quantization, or approximation, reaching a trade-off between accuracy and performance [19].

Federative approaches can also be applied, allowing not only the reduction of latency but also improving data privacy and strain on cloud networks by processing data locally [20]. More complex resource-demanding models can be deployed at the cloud, while being trained in a federated manner based on less complex models deployed at the edge. This allows the classification of labels directly at the edge and also at the cloud, adding another layer of elasticity to the model. Again, the miniaturization of the models such that the accuracy of a subset of labels remains high and the decision on when to execute the classification in the cloud remains an open research question. FL is a decentralized approach that enables distributing the load among edge nodes to train a collaborative model autonomously while keeping data localized [21] [22]. The edge devices share model updates to the cloud instead of raw data. Fig. 3 shows an example, depicting different ML models properly placed along the cloud-to-edge continuum to perform cognitive actions in order to improve the QoE of users. The different shapes represent the models that can be grouped across the continuum to be able to achieve more complex results.
FL has the following characteristics [23]:

- *Universality for cross-organizational scenarios,* allowing the building of a joint model from collaborators while maintaining underlying data locally;
- *Massively non-identically independent distribution,* keeping data widespread between edge nodes;
- *Decentralized technology,* enabling each client to influence the central model without having the need for a definitive central node; and

- *Equality of status for each node,* achieving a completely collaborative environment, where all parts enjoy equal status.

FL can also be applied to help train single models using diverse data sources from heterogeneous devices (such as those in the network access level) increasing the accuracy of the models independently of their data source setup. The elastic adaptation of FL systems to update partnerships considering the satisfaction of collaboration criteria (e.g., the minimum number of training partners) and the optimization of model accuracy is another open challenge to explore [12].

Cross-silo FL can help network operators collaborate on CCP services without sharing raw data. Using this approach, instead of centralizing data in one location, the model training is distributed across multiple devices (cross-device FL), services, or data spaces (silos), each holding local data. Thus, this approach allows training a global CCP module by exchanging relevant updates and aggregated insights [24] [25].

Another possibility that can be applied in edge and IoT scenarios is inference [26], which refers to applying a trained model on a new data sample to make a prediction. One main advantage of inference tasks is that they require fewer computational resources than training. Edge inference with Deep Neural Networks (DNNs) is a technique that will be explored. A common approach in distributed inference across mobile devices and edge servers is partitioning a pretrained DNN baseline between the devices and the edge server depending on the former's computational capabilities [27].

5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

As more and more data is generated in different fields, AI and ML are becoming popular, or rather, essential to help process and analyze all this information. One particular field that is enhanced by AI and ML is related to cognitive customer profiling, which aids the customer experience by tailoring the way businesses interact with their customers leveraging knowledge about their preferences and consumer patterns. Several data sources are combined to create accurate profiles, generating copious amounts of data that are usually handled in powerful data centers, commonly located in the cloud. These traditional cloud-based approaches bring three key challenges: additional impact on the communication network, increased latency, and privacy implications.

A novel paradigm called *Edge intelligence* has benefited from the booming of AI and ML applications and services. AI and ML bring these abilities to the devices deployed at the edge of the network infrastructure, improving the response time, reducing the network traffic by avoiding sending bulks of data to the network core, and simultaneously improving data privacy [28]. Federative approaches stand out as a way of distributing the load among the edge nodes, taking advantage of a distributed approach. Furthermore, recent developments in lightweight and miniaturized models, which can be executed in devices with low computing power such as those at the edge of the network, could improve the performance of a distributed CCP solution, overcoming the challenges previously identified.

ACKNOWLEDGEMENT

This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020; and by the project POWER (grant number POCI-01-0247-FEDER-070365), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020).

REFERENCES

[1] Chuntao Ding, Ao Zhou, Xiao Ma, and Shangguang Wang. "Cognitive Service in Mobile Edge Computing". In: *2020 IEEE International Conference on Web Services (ICWS)*. Beijing, China: IEEE, Oct. 2020, pp. 181–188. ISBN: 978-1-7281-8786-0. DOI: 10.1109/ICWS49710.2020.00031.

[2] Viktor H. Koelzer, Korsuk Sirinukunwattana, Jens Rittscher, and Kirsten D. Mertz. "Precision immunoprofiling by image analysis and artificial intelligence". In: *Virchows Arch* 474 (2019), pp. 511–522. DOI: https://doi.org/10.1007/s00428-018-2485-z.

[3] Claudia Gonzalez Viejo, Sigfredo Fuentes, Amruta Godbole, Bryce Widdicombe, and Ranjith R Unnithan. "Development of a low-cost e-nose to assess aroma profiles: An artificial intelligence application to assess beer quality". In: *Sensors and Actuators B: Chemical* 308 (2020), p. 127688. ISSN: 0925-4005. DOI: https://doi.org/10.1016/j.snb.2020.127688.

[4] Adrian Micu, Alexandru Capatina, Dragos Sebastian Cristea, Dan Munteanu, Angela-Eliza Micu, and Daniela Ancuta Sarpe. "Assessing an on-site customer profiling and hyper-personalization system prototype based on a deep learning approach". In: *Technological Forecasting and Social Change* 174.1 (Oct. 2021), p. 121289. ISSN: 0040-1625.

[5] Nisreen Ameen, Ali Tarhini, Alexander Reppel, and Amitabh Anand. "Customer experiences in the age of artificial intelligence". In: *Computers in Human Behavior* 114 (2021), p. 106548. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2020.106548.

[6] Young-Gab Kim Sang-Min Park. "User Profile System Based on Sentiment Analysis for Mobile Edge Computing". In: *Computers, Materials & Continua* 62.2 (May 2020), pp. 569–590. ISSN: 1546-2226.

[7] Ali Sunyaev. "Cloud Computing". In: *Internet Computing: Principles of Distributed Systems and Emerging Internet-Based Technologies*. Cham: Springer International Publishing, 2020, pp. 195–236. ISBN: 978-3-030-34957-8. DOI: 10.1007/978-3-030-34957-8_7.

[8] Keyan Cao, Yefan Liu, Gongjie Meng, and Qimeng Sun. "An Overview on Edge Computing Research". In: *IEEE Access* 8 (2020), pp. 85714–85728. DOI: 10.1109/ACCESS.2020.2991734.

[9] Shuiguang Deng, Hailiang Zhao, Weijia Fang, Jianwei Yin, Schahram Dustdar, and Albert Y. Zomaya. "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence". In: *IEEE Internet of Things Journal* 7.8 (2020), pp. 7457–7469. DOI: 10.1109/JIOT.2020.2984887.

[10] Cisco. *Cisco Annual Internet Report - White Paper*. [Online; accessed 01-February-2023]. 2020.

[11] Yaqiong Liu, Mugen Peng, Guochu Shou, Yudong Chen, and Siyu Chen. "Toward Edge Intelligence: Multiaccess Edge Computing for 5G and Internet of Things". In: *IEEE Internet of Things Journal* 7.8 (2020), pp. 6722–6747. DOI: 10.1109/JIOT.2020.3004500.

[12] Aaron Yi Ding, Ella Peltonen, Tobias Meuser, Atakan Aral, Christian Becker, Schahram Dustdar, Thomas Hiessl, Dieter Kranzlmüller, Madhusanka Liyanage, Setareh Maghsudi, Nitinder Mohan, Jörg Ott, Jan S. Rellermeyer, Stefan Schulte, Henning Schulzrinne, Gürkan Solmaz, Sasu Tarkoma, Blesson Varghese, and Lars Wolf. "Roadmap for Edge AI: A Dagstuhl Perspective". In: *SIGCOMM Computer Communication Review* 52.1 (Mar. 2022), pp. 28–33. ISSN: 0146-4833. DOI: 10.1145/3523230.3523235.

[13] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. "Train big, then compress: Rethinking model size for efficient training and inference of transformers". In: *37th International Conference on Machine Learning*. Online Conference: MLResearchPress, July 2020, pp. 5958–5968.

[14] Qian Zhang, Jie Lu, and Yaochu Jin. "Artificial intelligence in recommender systems". In: *Complex & Intelligent Systems* 7 (2021), pp. 439–457.

[15] Bogdan Walek and Petr Fajmon. "A hybrid recommender system for an online store using a fuzzy expert system". In: *Expert Systems with Applications* 212 (2023), p. 118565. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.118565.

[16] R.V. Karthik and Sannasi Ganapathy. "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce". In: *Applied Soft Computing* 108 (2021), p. 107396. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2021.107396.

[17] Jinke Ren, Guanding Yu, Yinghui He, and Geoffrey Ye Li. "Collaborative Cloud and Edge Computing for Latency Minimization". In: *IEEE Transactions on Vehicular Technology* 68.5 (2019), pp. 5031–5044. DOI: 10.1109/TVT.2019.2904244.

[18] Batyr Charyyev, Engin Arslan, and Mehmet Hadi Gunes. "Latency Comparison of Cloud Datacenters and Edge Servers". In: *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. 2020, pp. 1–6. DOI: 10.1109/GLOBECOM42002.2020.9322406.

[19] Umar Ibrahim Minhas, JunKyu Lee, Lev Mukhanov, Georgios Karakonstantis, Hans Vandierendonck, and Roger Woods. "Increased Leverage of Transprecision Computing for Machine Vision Applications at the Edge". In: *Journal of Signal Processing Systems* 1.94 (Oct. 2022), pp. 1101–1118. ISSN: 19398115.

[20] Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. "Federated learning review: Fundamentals, enabling technologies, and future applications". In: *Information Processing & Management* 59.6 (Nov. 2022), pp. 1–24. ISSN: 0306-4573.

[21] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. "Federated Learning in Mobile Edge Networks: A Comprehensive Survey". In: *IEEE Communications Surveys & Tutorials* 22.3 (2020), pp. 2031–2063. DOI: 10.1109/COMST.2020.2986024.

[22] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. "A survey on federated learning". In: *Knowledge-Based Systems* 216 (2021), p. 106775. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2021.106775.

[23] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. "A review of applications in federated learning". In: *Computers & Industrial Engineering* 149 (2020), p. 106854. ISSN: 0360-8352. DOI: https://doi.org/10.1016/j.cie.2020.106854.

[24] Saikishore Kalloori and Abhishek Srivastava. "Towards cross-silo federated learning for corporate organizations". In: *Knowledge-Based Systems* 289.1 (Apr. 2024), pp. 1–11. ISSN: 0950-7051.

[25] Jianfeng Lu, Bangqi Pan, Juan Yu, Wenchao Jiang, Jianmin Han, and Zhiwei Ye. "Towards energy-efficient and time-sensitive task assignment in cross-silo federated learning". In: *Journal of King Saud University - Computer and Information Sciences* 35.4 (Apr. 2023), pp. 63–74. ISSN: 1319-1578.

[26] Carole-Jean Wu, David Brooks, Kevin Chen, Douglas Chen, Sy Choudhury, Marat Dukhan, Kim Hazelwood, Eldad Isaac, Yangqing Jia, Bill Jia, Tommer Leyvand, Hao Lu, Yang Lu, Lin Qiao, Brandon Reagen, Joe Spisak, Fei Sun, Andrew Tulloch, Peter Vajda, Xiaodong Wang, Yanghan Wang, Bram Wasti, Yiming Wu, Ran Xian, Sungjoo Yoo, and Peizhao Zhang. "Machine Learning at Facebook: Understanding Inference at the Edge". In: *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 2019, pp. 331–344. DOI: 10.1109/HPCA.2019.00048.

[27] S. Joe Qin and Leo H. Chiang. "Advances and opportunities in machine learning for process data analytics". In: *Computers & Chemical Engineering* 126.1 (July 2019), pp. 465–473. ISSN: 0098-1354.

[28] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing". In: *Proceedings of the IEEE* 107.8 (June 2019), pp. 1738–1762. ISSN: 1558-2256.

AUTHORS

Karima Velasquez got her Ph.D. in information science and technologies from the University of Coimbra in 2021. Since 2014 she has been working as a researcher attached to the Laboratory of Communications and Telematics at the Centre for Informatics and Systems of the University of Coimbra, where

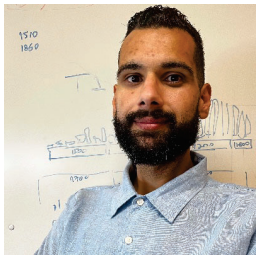
she also works as an assistant professor in the Department of Informatics Engineering. She has over 15 years of experience working on research groups and participating in national and international research projects. She has co-authored over 30 conference and journal papers and is a member of the Programme Committee of WMCN, LANC, FlexNGIA, DRCN, and DML-ICC. She is an assistant editor for the Elsevier Internet of Things journal. Her research interests include cloud-to-edge continuum, edge intelligence, and network management and orchestration.

David Perez Abreu received his Ph.D. in information science and technologies from the University of Coimbra in 2021. He has published several conference and journal papers and has actively participated in different national and international research projects. He is part of the program committee of NOMS, SLICO - WoWMoM, LCN, ISCC, and DRCN. His research interests include cloud-to-edge computing, softwarized networks, and zero-touch network and service management. He worked as a researcher in the Laboratory for Mobile and Wireless Networks at the Central University of Venezuela from 2006 until 2014. He is an invited assistant professor at the Department of Informatics Engineering at the University of Coimbra and a senior researcher at the Laboratory for Informatics and Systems at the Pedro Nunes Institute.

Pedro Neves received his M.Sc. and Ph.D. degrees in electronics and telecommunications engineering from the University of Aveiro, Portugal, in 2006 and 2012 respectively. After graduation, he became a research fellow at the Telecommunications Institute, where he worked on European-funded projects on broadband wireless access networks. In June 2006 he joined PT Inovação, working on heterogeneous wireless environments in the context of European and Eurescom-funded projects. He participated in more than 10 international collaborative projects, is co-author of six international books, and has published more than 30 articles in journals and conference proceedings.

Miguel Santos has a degree in electrical and computer engineering from the University of Oporto (FEUP), Oporto, Portugal, and a Master of Science degree in electrical and computer engineering from the University of Florida, Florida, United States of America. As part of PT Inovação/Altice Labs since 2002 he has worked mainly in data network services and control areas. It was part of the development of several Altice Labs products from which we can highlight charging control and policy, and also guest Wi-Fi management. Currently, he works as leader of the product management team of network and service platforms area at Altice Labs. His research interests are in the following technologies: 5G, NFV, SDN, and AI.

Fernando Santiago got his bachelor's degree in electronics and telecommunications engineering from the University of Aveiro, Portugal, in 1993. He is currently a product manager at Altice Labs. With more than 25 years of experience in the telecommunications industry, he has had several positions ranging from software developer, team leader, project manager, head of the unit, product manager, and product owner. He has worked on several projects related to audio compression technologies, enterprise applications, e-health /telemedicine, M2M/IoT, Martech, and data/AI platforms. He has co-authored several articles on various topics in his fields of work.



Nuno Lourenço is an assistant professor at the Department of Informatics Engineering of the University of Coimbra, where he obtained his PhD in information science and technology in 2016. He is the current coordinator of the Evolutionary and Complex Systems (ECOS)

group and has been a member of the Centre for Informatics and Systems of the University of Coimbra (CISUC) since 2009. Formerly, he was appointed as a senior research officer at the University of Essex in the United Kingdom. His main research interests are in the areas of bio-inspired algorithms, optimization, and machine learning.. He is the co-creator of structured grammatical evolution and probabilistic grammatical evolution, and DENSER, a novel approach to automatically design deep artificial neural networks using evolutionary computation. He served as chair in the main conferences of the evolutionary computation field, namely EuroGP and PPSN. He is a member of the Programme Committee of GECCO, PPSN, and EuroGP; and an executive board member of SPECIES. He has authored or co-authored more than 60 articles in journals and top conferences from the evolutionary computation and artificial intelligence areas and he has been involved as a researcher in 13 projects (national and international).

Marilia Curado is a full professor at the Department of Informatics Engineering of the University of Coimbra, Portugal, where she obtained her Ph.D. in computer engineering in 2005. She is currently Director of the Laboratory for Informatics and Systems of Pedro Nunes Institute. Her research interests include Resilience and quality of service in computer networks, energy efficient communications, the Internet of Things, and communications in the cloud. She co-authored over 200 publications including more than 40 ISI journals. She is a member of the editorial board of Elsevier Computer Networks, Elsevier Computer Communications, Elsevier Internet of Things, and Wiley Internet Technology Letters and has been involved in the scientific organization and coordination of several international conferences. She has participated in many national and international projects.

Edmundo Monteiro is a full professor at the University of Coimbra (UC), Portugal, from where he graduated in electrical engineering (informatics speciality) in 1984, received his Ph.D. in electrical engineering (computer communications), and the Habilitation in informatics engineering, in 1996 and 2007 respectively. He is currently Head of the Informatics Engineering Department of the University of Coimbra. He is a senior researcher at the Centre for Informatics and Systems of the University of Coimbra (CISUC). He has more than 35 years of research and industry experience in the fields o f C omputer c ommunications, w ireless and mobile communications, Quality of service, network and service management, cybersecurity, critical infrastructure protection, cloud networking, Internet of Things, and sensor networks. His publications include six books (authored and edited), several book chapters, and over 200 papers in international refereed journals and conferences. He is also co-author of nine international patents. He is a member of the editorial board of Springer Wireless Networks and ITU Journal on Future and Evolving Technologies journals. Edmundo Monteiro is member of Ordem dos Engenheiros (the Portuguese Engineering Association), and a senior member of the IEEE Communication Society, IEEE Computer Society, and ACM SIGCOMM. He is also the Portuguese representative in IFIP TC6 (Communication Systems).

Carlos Amaro got his Bachelor's degree in computer science and engineering systems from the University of Minho, Portugal. He is currently a project manager in Altice Labs, PT, with over ten years of experience in IT and Telco projects. His specialities include project management, mainly in IT and telecommunications. He is PMP certified. His current focus is on intelligent networks.