Towards Cognitive Service Delivery on B5G through AIaaS Architecture

Larissa Ferreira Rodrigues Moreira[1,2], **Rodrigo Moreira**[1], **Flávio de Oliveira Silva**[2,3], **André Ricardo Backes**[4]

[1]Federal University of Viçosa (UFV)
Rio Paranaíba – MG – Brazil

[2]Faculty of Computing (FACOM) – Federal University of Uberlândia (UFU)
Uberlândia – MG – Brazil

[3]Department of Informatics – School of Engineering
University of Minho – Braga, Portugal

[4]Department of Computing (DC) – Federal University of São Carlos (UFSCar)
São Carlos – SP – Brazil

{larissa.f.rodrigues, rodrigo}@ufv.br, flavio@di.uminho.pt,

{larissarodrigues, flavio}@ufu.br, arbackes@yahoo.com.br

Abstract. *Artificial Intelligence (AI) is pivotal in advancing mobile network systems by facilitating smart capabilities and automation. The transition from 4G to 5G has substantial implications for AI in consolidating a network predominantly geared towards business verticals. In this context, 3GPP has specified and introduced the Network Data Analytics Function (NWDAF) entity at the network's core to provide insights based on AI algorithms to benefit network orchestration. This paper proposes a framework for evolving NWDAF that presents the interfaces necessary to further empower the core network with AI capabilities B5G and 6G. In addition, we identify a set of research directions for realizing a distributed e-NWDAF.*

1. Introduction

To support different industry verticals, such as the Internet of Everything (IoE), Smart Factory, or Self Driving, many challenges and stringent requirements have been imposed on the standardization and development processes of 5G mobile networks [Coronado et al. 2022]. In particular, numerous advances have been made in the control plane of this network, such as network slicing [Moreira and Silva 2021], beamforming [Brilhante et al. 2023], closed-loop automation, Artificial Intelligence (AI), and security enforcement [Silva et al. 2023] to support improved connectivity and chaining of virtualized network functions.

Different AI techniques were applied to the core of the 5G network, aiming for efficient resource management and the delivery of connectivity to application vertices. Some disruptive techniques are built apart from the 3rd Generation Partnership Project (3GPP) specifications, while others are aligned with 3GPP, which is the case of the Network Data Analytics Function (NWDAF) entity idealized in Release 15 [3GPP 2023]. The role of NWDAF is to collect information from various network functions (NFs)

and to support AI-based insights for the benefit of network management and orchestration [Ishteyaq et al. 2024]. However, the quality of the NWDAF operation comes across all fields related to the availability of high-quality data; that is, the evolution of models depends on the data generated in the operator's domain, making generalization difficult [Gkonis et al. 2024].

This paper aimed to design an evolution proposal aligned with 3GPP for NWDAF, an analytics entity of a mobile 5G network core. The underlying concept behind our Evolved NWDAF (*e*-NWDAF) is that it enables different players to access or offer third-party cognitive services in a mobile network through a northern interface on NWDAF. In our vision, we present the interfaces, relationships, and behaviors for the realization of the evolved *e*–NWDAF.

In addition, we discuss the AI for Networking (AI4Net) paradigms as a way to enhance the NWDAF functionalities of 3GPP and propose Networking for AI (Net4AI) to deliver cognitive services at the core of mobile networks through the Artificial Intelligence as a Service (AIaaS) [Rodrigues Moreira et al. 2023] or NetApps [Slamnik-Kriještorac et al. 2022] architectures. Our main contribution lies in an architectural framework to evolve NWDAF to support different paradigms of offering cognitive services, as well as presenting a short survey with research directions towards evolving NWDAF.

The remainder of this paper is organized as follows. Section 2 surveys the related work, and Section 3 presents the rationale of our approach. In Section 4, we summarize a set of search directions around NWDAF and present conclusions in Section 5.

2. Related Work

This section provides a comparative review of state-of-the-art solutions that relate to both aspects of our proposal. The first refers to generic architectures for delivering cognitive services, while the other specifies approaches that target advancements in the core of the B5G network to enable AI verticals.

2.1. Generic AI Architectures

[Shah et al. 2022] proposed incorporating AIaaS, Service as a Service (SaaS) and Infrastructure as a Service (IaaS) to curb the spread of unethical content on social media platforms. They proposed using AIaaS to identify and remove such content, which was classified into immoral, cyberbullying, and dislike categories. The authors employed traditional supervised learning algorithms but did not evaluate the effectiveness of deep learning techniques despite claiming that their approach is suitable for large datasets.

[Fortuna et al. 2023] investigated the potential benefits of using on-premise AIaaS solutions for small and medium-sized businesses. The study evaluated the capabilities of AIaaS and explored the feasibility of implementing it using open-source, user-friendly technologies. These technologies allow businesses to control their data, processes, and costs while minimizing reliance on third-party vendors or vendor lock-in risks.

[Guntupalli and Rudramalla 2023] introduced a method to guarantee data compression, integrity, and confidentiality in AIaaS deployments, which are online repositories that provide various Machine Learning (ML) services and tools to users. However,

selecting the best machine learning model or AI service can be difficult. To address this challenge, [Cerar and Hribar 2023] proposed a dynamic approach that utilizes Deep Reinforcement Learning (DRL) to select the most appropriate ML model and evaluate the method for a specific case of energy consumption forecasting.

[Zhang et al. 2023] proposed an AIaaS model deployment architecture for edge computing that enables the configuration of data quality and model complexity ratios for AI tasks, providing low-latency AI capabilities. [Hajipour et al. 2023] proposed a business plan incorporating AI products and services through AIaaS, featuring a strategic approach, step-by-step plan, and heuristic pricing model. However, the plan has not yet been implemented.

[Baccour et al. 2023] proposed a novel platform architecture for deploying zero-touch Pervasive Artificial Intelligence as a Service (PAIaaS) in 6G networks, supported by a blockchain-based smart system. They tested a use case for Federated Learning (FL), where the service consumer deployed distributed training on the 6G infrastructure with the assistance of the service provider's application, utilizing widespread devices.

Table 1 summarizes the related works described above and highlights the differences demonstrating how our proposal represents a significant advancement in the field. The "Native API for Edge" column characterizes the solutions to offer native API in edge computing applications. The "Multiple AI Facilities" column characterizes the solutions to offer multiple AI facilities such as different ML algorithms, optimization methods, and feature extraction. Finally, the "6G enabled" column refers to 6G-enabled solutions.

Table 1. Short State-of-the-Art Survey.

Approach	Native API for Edge	Multiple AI Facilities	6G enabled
[Shah et al. 2022]	○	●	○
[Fortuna et al. 2023]	○	●	○
[Guntupalli and Rudramalla 2023]	○	●	○
[Cerar and Hribar 2023]	○	●	○
[Zhang et al. 2023]	○	●	○
[Hajipour et al. 2023]	○	●	○
[Baccour et al. 2023]	○	○	●
Our approach	●	●	●

2.2. B5G-Enabled AI Architectures

[Li et al. 2021] proposed a cognitive service architecture for the 6G core network, inspired by the nervous system of an octopus, to enhance network performance and meet high service quality requirements. The cognitive service architecture aims to provide real-time perception, AI reasoning abilities, and a knowledge graph management system to adjust the core network dynamically. A case study is conducted to evaluate the performance of the cognitive service architecture, demonstrating significant time savings in session establishment compared to noncognitive service architecture.

[Manias et al. 2022] developed a functional prototype of NWDAF within a 5GC network, using unsupervised learning and clustering techniques to analyze intra-network interactions and optimize operations. The innovation lies in creating a working prototype for the 5GC network and NWDAF, enabling practical data-driven techniques and offering insights for future research and network management.

[Fiore 2023] introduced a full network sensing paradigm, which systematically integrates sensing functionalities across all mobile network architecture domains. Their approach aims to unlock mobile networks' untapped potential, allowing services to access rich metadata for knowledge discovery and informed policy-making. The authors highlight the need for a major shift in designing mobile networks for sensing, emphasizing the creation of new markets and innovative services that benefit society.

[Hossain et al. 2023] introduced a distributed ML approach using NWDAF to optimize network operations. They presented two variations: one based on federated learning and the other on split learning, aiming to enhance data security, accuracy, and automation. The proposed architectures can greatly contribute to automating operations in 5G+ networks, yet challenges remain to be addressed.

[Jeon and Pack 2024] introduced a hierarchical network data analytics framework (H-NDAF), which distributes inference tasks to multiple leaf NWDAFs and centralizes training tasks at the root NWDAF, ensuring timely inference results while maintaining high accuracy. The H-NDAF was implemented using open-source software (free5GC) and evaluated through extensive simulations, particularly optimizing the policy for User Equipment (UE) data flows.

3. Proposed Core Enhancement

This paper proposes a background for NWDAF enhancement Beyond 5G or 6G networks. The fundamental idea behind this enhancement is to extend the functions that 3GPP already designed in Release 18 for NWDAF from two perspectives. The first perspective is to allow the network to benefit from AI capabilities in a more sophisticated manner, with the operation being of the AI for Networking (AI4Net) type. The second perspective is the return to enabling AI functions for third parties through mobile networks, the operation being of the Networking for AI (Net4AI) type.

The current standardization of NWDAF provides analytics functions for mobile networks through a set of messages for data collection, reporting, subscription, and exposure of core entity metrics. NWDAF allows for updating the pipeline with new analytics types, algorithms, or models and integrating them into operations. However, the NWDAF standardization model is managed solely by network operators, hindering third-party innovation in delivering cognitive services. To address this, we propose combining and evolving NWDAF with AIaaS and NetApp concepts, currently in development, to specify how third-party services can be integrated and specified within the network framework.

AIaaS refers to an architectural framework for lifecycle management and delivery of AI services that aims to seamlessly deploy and embed intelligence capabilities to applications or edge devices [Rodrigues Moreira et al. 2023]. In line with European Telecommunications Standards Institute (ETSI), the concept of NetApps was created, which is a 5G-enabled virtual application, and an evolution of the concept of Virtualized Network Function (VNF), which provides functionalities to deliver complex services [Slamnik-Kriještorac et al. 2022]. Thus, in Figure 1, we present the intervention proposed in this paper, which we call evolved NWDAF (*e*-NWDAF).

In Figure 1, we position the main elements of the 5G core, and on the left, we highlight *e*-NWDAF with its additional functions: AI4Net and Net4AI. In our *e*-NWDAF, we

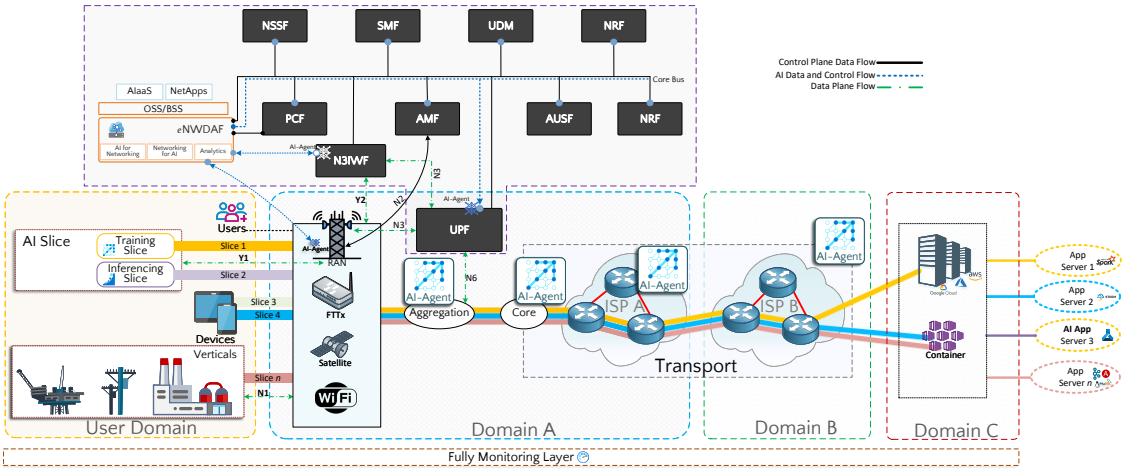

Figure 1. Evolved NWDAF to support AIaaS Cognitive Service Delivering.

idealize two evolution perspectives, namely, the Operations Support System and Business Support System (OSS/BSS) interface, to enable the consumption and delivery of third-party cognitive services to users. The second refers to an architectural deployment model using AI-Agents as microservices composed and distributed to each specific function of the network core.

The evolution perspective of NWDAF aims to enable "Training Slice" or "Inferencing Slice" for users [Shen et al. 2022]. In addition to the potential to benefit network management at the core of the Service Provider, decoupling NWDAF functionalities exclusively to the network operator and enabling the wake of new applications, such as Mobile Network Metrics as a Service. The second perspective of evolution enables a refined closed loop of AI capabilities available to each core entity in the network.

With *e*-NWDAF, we envision performing network slices beyond connectivity, that is, network slices for training AI or inference models. When idealizing 5G, a set of network requirements and business verticals would be enabled on mobile networks such as Ultra Reliable Low Latency Communications (URLLC), Massive Machine Type Communications (mMTC), or enhanced Mobile Broadband (eMBB). For each of these verticals, the technical challenges of network slices as well as their Key Performance Indicators (KPIs) became known. With *e*-NWDAF, we envision a new vertical format for applications that can be enabled in B5G networks, such as NetAPPs or AIaaS facilities.

4. Research Directions

Towards distributed NWDAF boosted by AI, we conducted a short state-of-the-art analysis on the three main indexing databases, Association for Computing Machinery (ACM), ScienceDirect, and Institute of Electrical and Electronics Engineers (IEEE) Xplore, using the search string as below. The word combination provides a focused yet comprehensive approach to research, bridging network data analytics with advanced distributed learning techniques. This intersection aims brings practical challenges in data analysis and offers theoretical grounds to evolve AI service composition.

```
("NWDAF") AND ("Federated Learning")
```

We defined this search string because we are looking for approaches that assume that analytics functionalities at the core of the network can use distributed learning approaches such as federated learning. Analyzing these 35 papers, we sought to highlight and summarize the challenges and possible directions based on the notes of these authors. We only considered papers that contained consistent and practical approaches to NWDAF and excluded short papers or surveys. In Table 2, we summarize these challenges and highlight the research directions that should be addressed to realize or evolve the embedding intelligence in B5G systems.

Table 2. Challenges and Possible Directions towards Realization and Evolving of NWDAF.

Challenge	Possible Directions
Practical Deployment	• Efficient management and orchestration for space-aerial-terrestrial-ocean integrated. • Intelligent and flexible management approaches for NWDAF. • Model and feature selection. • Training acceleration. • Containerization and interface design. • Self-Coordination and self-organization of NWDAF leafs. • Cross-domain training collaboration. • Addressing statistical heterogeneity. • AI/ML model maintenance.
Scalability	• Interdomain scalability. • Decentralization of NWDAF for real-time applications.
Privacy and Security	• Ensuring security and privacy protection on NWDAF operation. • Use FL to enable collaborative model training without disclosing private raw data. • Trustworthiness of AI in NWDAF. • Security Overhead on Training and Inference life-cycle. • Developing robust security measures that protect against intrusions
AI Technologies	• AI/ML integration. • Integrated FL services. • Address the variance in FL update to reduce the effects of stragglers. • Cooperation and Sharing among multiple NWDAFs to train an ideal ML model.
Communication	• Efficient protocols for NWDAF message exchanging. • Enhancing FL methods to minimize communication overhead.
AI Explainability	• Determining the reason of optimal placement of ML models on NWDAF leafs.
Standardization	• 3GPP and Open Radio Access Network (O-RAN) for AI/ML. • Interoperability with other systems (e.g. *non*-3GPP systems).

5. Concluding Remarks

The present paper offers an evolutionary proposal for the NWDAF system, which adopts a distributed architecture and employs AI-Agents to interface with the various core functions of the mobile network. We undertook a comprehensive analysis of the core functions, identifying the aspects that are most suitable for enhancement following the specifications outlined by 3GPP. While NWDAF in its original form represents a significant step forward in the field, it can be further refined to enable a broader range of benefits, particularly in cognitive service delivery.

We designed interfaces that enable third parties to receive cognitive services through Net4AI and expanded the fundamental principles of NWDAF to encompass third-party cognitive services in the context of AI4Net. We identified research directions that focused on creating and verifying *e*-NWDAF architectures for B5G deployment. One of the limitations of this paper is the need for empirical validation of theoretical concepts. In future work, we plan to address this limitation by presenting validation and benchmark results.

Acknowledgments

We thank CAPES and FAPESP MCTIC/CGI Research project 2018/23097-3 for the financial support. André R. Backes gratefully acknowledges the financial support of CNPq (National Council for Scientific and Technological Development, Brazil) (Grant #307100/2021-9). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This work was also partially supported by RNP with resources from MCTIC, Grant No. 01245.010604/2020-14, under the Brazil 6G Project and Centro ALGORITMI, funded by Fundação para a Ciência e Tecnologia (FCT) within the RD Units Project Scope 2020-2023 (UIDB/00319/2020) for partially support this work.

References

3GPP (2023). 5G System; Network Data Analytics Services; Stage 3. Technical Report TR 29.520 V15.3.0 (2019-04), 3GPP.

Baccour, E., Allahham, M. S., Erbad, A., Mohamed, A., Hussein, A. R., and Hamdi, M. (2023). Zero Touch Realization of Pervasive Artificial Intelligence as a Service in 6G Networks. *IEEE Communications Magazine*, 61(2):110–116.

Brilhante, D. d. S., Manjarres, J. C., Moreira, R., de Oliveira Veiga, L., de Rezende, J. F., Müller, F., Klautau, A., Leonel Mendes, L., and P. de Figueiredo, F. A. (2023). A Literature Survey on AI-Aided Beamforming and Beam Management for 5G and 6G Systems. *Sensors*, 23(9).

Cerar, G. and Hribar, J. (2023). Machine Learning Operations Model Store: Optimizing Model Selection for AI as a Service. In *2023 International Balkan Conference on Communications and Networking (BalkanCom)*, pages 1–5, İstanbul, Turkiye. IEEE.

Coronado, E., Behravesh, R., Subramanya, T., Fernàndez-Fernàndez, A., Siddiqui, M. S., Costa-Pérez, X., and Riggio, R. (2022). Zero touch management: A survey of network automation solutions for 5g and 6g networks. *IEEE Communications Surveys & Tutorials*, 24(4):2535–2578.

Fiore, M. (2023). Full Network Sensing: Architecting 6G Beyond Communications. *IEEE Network*, 37(3):232–239.

Fortuna, C., Mušić, D., Cerar, G., Čampa, A., Kapsalis, P., and Mohorčič, M. (2023). *On-Premise Artificial Intelligence as a Service for Small and Medium Size Setups*. Springer International Publishing, Cham.

Gkonis, P. K., Nomikos, N., Trakadas, P., Sarakis, L., Xylouris, G., Masip-Bruin, X., and Martrat, J. (2024). Leveraging network data analytics function and machine learning for data collection, resource optimization, security and privacy in 6g networks. *IEEE Access*, 12:21320–21336.

Guntupalli, N. and Rudramalla, V. (2023). Artificial intelligence as a service: Providing integrity and confidentiality. In Morusupalli, R., Dandibhotla, T. S., Atluri, V. V., Windridge, D., Lingras, P., and Komati, V. R., editors, *Multi-disciplinary Trends in Artificial Intelligence*, pages 309–315, Cham. Springer Nature Switzerland.

Hajipour, V., Hekmat, S., and Amini, M. (2023). A value-oriented Artificial Intelligence-as-a-Service business plan using integrated tools and services. *Decision Analytics Journal*, page 100302.

Hossain, M. A., Hossain, A. R., Liu, W., Ansari, N., Kiani, A., and Saboorian, T. (2023). A Distributed Collaborative Learning Approach in 5G+ Core Networks. *IEEE Network*, pages 1–8.

Ishteyaq, I., Muzaffar, K., Shafi, N., and Alathbah, M. A. (2024). Unleashing the Power of Tomorrow: Exploration of Next Frontier With 6G Networks and Cutting Edge Technologies. *IEEE Access*, 12:29445–29463.

Jeon, Y. and Pack, S. (2024). Hierarchical Network Data Analytics Framework for 6G Network Automation: Design and Implementation. *IEEE Internet Computing*, pages 1–9.

Li, Y., Huang, J., Sun, Q., Sun, T., and Wang, S. (2021). Cognitive service architecture for 6g core network. *IEEE Transactions on Industrial Informatics*, 17(10):7193–7203.

Manias, D. M., Chouman, A., and Shami, A. (2022). An NWDAF Approach to 5G Core Network Signaling Traffic: Analysis and Characterization. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 6001–6006.

Moreira, R. and Silva, F. (2021). Towards 6G Network Slicing. In *Anais do I Workshop de Redes 6G*, pages 25–30, Porto Alegre, RS, Brasil. SBC.

Rodrigues Moreira, L. F., Moreira, R., Travençolo, B. A. N., and Backes, A. R. (2023). An artificial intelligence-as-a-service architecture for deep learning model embodiment on low-cost devices: A case study of covid-19 diagnosis. *Applied Soft Computing*, 134:110014.

Shah, F., Anwar, A., ul haq, I., AlSalman, H., Hussain, S., and Al-Hadhrami, S. (2022). Artificial Intelligence as a Service for Immoral Content Detection and Eradication. *Scientific Programming*, 2022:6825228.

Shen, X., Gao, J., Wu, W., Li, M., Zhou, C., and Zhuang, W. (2022). Holistic network virtualization and pervasive network intelligence for 6g. *IEEE Communications Surveys & Tutorials*, 24(1):1–30.

Silva, C., Cruz, A., Andrade, R., and Rodrigues, E. (2023). Requisitos e Desafios de Segurança e Privacidade em Redes 6G. In *Anais do III Workshop de Redes 6G*, pages 19–24, Porto Alegre, RS, Brasil. SBC.

Slamnik-Kriještorac, N., Landi, G., Brenes, J., Vulpe, A., Suciu, G., Carlan, V., Trichias, K., Kotinas, I., Municio, E., Ropodi, A., and Marquez-Barja, J. M. (2022). Network Applications (NetApps) as a 5G booster for Transport & Logistics (T&L) Services: The VITAL-5G approach. In *2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 279–284.

Zhang, W., Zeadally, S., Li, W., Zhang, H., Hou, J., and Leung, V. C. M. (2023). Edge AI as a Service: Configurable Model Deployment and Delay-Energy Optimization With Result Quality Constraints. *IEEE Transactions on Cloud Computing*, 11(2):1954–1969.