Artificial Intelligence as a Service Architecture: an innovative approach for Computer Vision applications

Larissa Ferreira Rodrigues Moreira[1,2],
Bruno Augusto Nassif Travençolo[1], **André Ricardo Backes**[3]

[1]School of Computer Science – Federal University of Uberlândia (UFU)
Uberlândia – MG – Brazil

[2]Institute of Exacts and Technological Sciences – Federal University of Viçosa (UFV)
Rio Paranaíba – MG – Brazil

[3]Departament of Computing – Federal University of São Carlos (UFSCar)
São Carlos – SP – Brazil

{larissarodrigues, travencolo}@ufu.br

larissa.f.rodrigues@ufv.br; arbackes@yahoo.com.br

Abstract. *In recent years, Artificial Intelligence (AI) has grown significantly across various domains such as transportation, healthcare, and security. However, current implementations of intelligent services face challenges in enhancing the personalized and large-scale use of AI. This work presents Artificial Intelligence as a Service (AIaaS), an innovative approach to effectively manage the lifecycle of diverse AI models and paradigms, while offering them as a service for heterogeneous devices, multiple users, and modern applications. We explored the feasibility of delivering AI resources in different network architectures, such as edge computing and mobile networks, which provide a flexible and scalable environment that allows users to acquire cognitive services in the AI life cycle. Also, our approach facilitates personalized and scalable AI solutions, fostering innovation and expediting the deployment of intelligent applications across diverse contexts, making it suitable for real-world scenarios.*

Ph.D. Thesis defended on September 23, 2024 (FACOM – UFU). Committee: Prof. André R. Backes (UFSCar), Prof. Bruno A. N. Travençolo (UFU), Prof. Alessandra A. Paulino (UFU), Prof. João Henrique S. Pereira (UFU), Prof. Dalcimar Casanova (UTFPR), and Prof. João B. Florindo (UNICAMP). Full text available at: doi.org/10.14393/ufu.te.2024.675.

1. Introduction

Artificial Intelligence (AI) has emerged as the most prevalent technology in recent decades, with projections of generating a $19.9 trillion economic impact by 2030. Gartner emphasizes AI-driven development in strategic technology trends, highlighting innovations such as cloud AI services, computer vision, edge AI, and model operationalization [Jaffri and Sicular 2023]. The expansion of computational services supports smart applications across various domains, including healthcare, elder care, entertainment, education, precision agriculture, and security [Ahmed et al. 2022, Rodrigues Moreira et al. 2025]. Despite the vast amounts of data generated and the need

for harmonized hardware, embedded systems enhance efficiency, with the global AI market expected to reach $22.4 billion by 2030 [Luxton 2023].

The deployment of context-specific AI services is essential, particularly for Machine Learning (ML) algorithms used in disease identification using images, which require substantial hardware resources for training, validation, testing, and deployment. This iterative process emphasizes the critical necessity of systematic AI artifact lifecycle management, supporting the consolidation of large-scale ecosystems across diverse contexts and applications [Ahmed et al. 2022].

1.1. Problem Statement

The advancement of AI methods combined with mobile and embedded devices has led to the development of intelligent devices, such as smartphones, smartwatches, and sensors, integrated into various appliances. However, there is a crucial need to explore the application of deep neural networks to these devices to support sensor inference, paving the way for the next generation of intelligent devices for future network applications [Ahmed et al. 2022]. Despite their potential benefits, the current AI resources lack organization and personalization, which hampers the robustness and efficiency of applications, particularly during software updates. Consequently, there is an opportunity to enhance the capabilities of application providers by outsourcing and selectively applying AI services, thereby improving the intelligent service deployment.

Moreover, delivering AI resources as a service provides a systematic approach for managing these artifacts and facilitating updates and adjustments without necessitating modifications to the application itself. This organizational method supports large-scale ecosystems across diverse contexts and applications. To support this thesis, we provide an overview of the current state of deep learning-based image classification on embedded devices (from 2013 to 2023), highlighting the trends, challenges, and advancements in this field [Rodrigues Moreira et al. 2025]. The dominance of low-cost devices, such as Raspberry Pi and ARM Cortex (Figure 1), owing to their affordability and robust community support, emphasizes the potential of deploying AI cognitive services on accessible hardware, which is essential for the development of Artificial Intelligence as a Service (AIaaS).

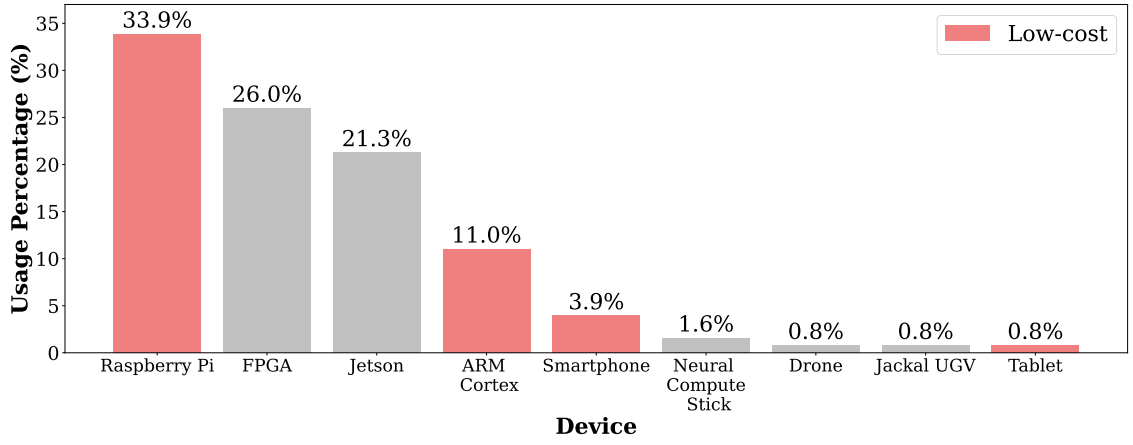

Figure 1. Device usage for deep learning. Low-cost devices in red.

Research Hypothesis. This work is based on the following hypothesis:

> **HYPOTHESIS:** *AI as a Service Architecture improves life cycle management in Computer Vision field, such as model training, validation, testing, and model fine-tuning management by providing to users and applications baseline infrastructure and methods to handle cognitive service delivery.*

1.2. Goals and Contributions

This research aims to develop an **Architecture for delivering AI as a Service**, providing on-demand solutions for managing ML models, fine-tuning services, and datasets. Key objectives include evaluating the approach through extensive image dataset experiments, proposing a new AIaaS architecture, mapping challenges in edge prediction, investigating hyperparameter adjustment in cloud computing, and exploring data augmentation techniques to address small and unbalanced datasets.

Our main contributions are:

- A new AIaaS Architecture to deliver AI as a Service and is dedicated to providing on-demand solutions for users and applications.
- Evaluation of the computational cost of embedding deep learning models in different devices.
- An analysis to assess the relationship between the computational cost of a Convolutional Neural Network (CNN) and the number of its parameters.
- AI service design different forms, being: ML model management; trained model tuning service; and dataset management.
- A proof of concept to showcase the effectiveness of our proposed architecture in delivering AI models in a novel way.

2. Related Work

This section provides a comprehensive survey of the state-of-the-art solutions in AIaaS. [Chan et al. 2013] and [Baldominos et al. 2014] introduced Hadoop-based platforms for PredictionIO and ad recommendations, respectively, though they are domain-specific. [Ribeiro et al. 2015] proposed a scalable Machine Learning as a Service (MLaaS) architecture for prediction tasks, while [Li et al. 2017] created a scalable MLaaS architecture for Uber, combining deep learning and traditional algorithms. Despite their comprehensiveness, they lack customization and support for low-cost devices. [Yao et al. 2022] developed VenusAI for large-scale scientific computing, but it has limitations in large-scale support.

AIaaS applications are diverse, from autonomous driving [De Caro et al. 2022] to content moderation in social networks [Shah et al. 2022]). Market analyses by [Lewicki et al. 2023] and on-premise solutions for small and medium-sized companies by [Fortuna et al. 2023] highlight potential biases and feasibility. Advanced techniques include data-integrity algorithms [Guntupalli and Rudramalla 2023], dynamic model selection [Cerar and Hribar 2023], and decentralized intelligence [Lomonaco et al. 2023]. [Zhang et al. 2023] proposed an AIaaS model for edge computing, while [Hajipour et al. 2023] offered a business plan for AIaaS commercialization.

Future technologies focus on B5G and 6G, with [Baccour et al. 2023] proposing zero-touch Pervasive Artificial Intelligence as a Service (PAIaaS) and [Oliveira et al. 2024a] and [Nadar and Härri 2024] integrating Network Data Analytics Function (NWDAF) with AIaaS for anomaly detection in 5G.

Companies such as Amazon, Google, and Microsoft have proprietary platforms that prevent external developers from adding specific functionalities. In response, our proposed AIaaS innovates by allowing inclusion, enhancement, and adaptation of new ML models for various applications. In addition, this work addresses these issues by evolving NWDAF with AIaaS to integrate third-party services, proposing a framework for AI service lifecycle management, and addressing the limitations of existing state-of-the-art solutions.

As summarized in Table 1, related works highlight the absence of key features such as native API support for smartphones and low-cost devices, dataset management, multiple AI facilities (e.g., ML algorithms, optimization methods, feature extraction), and computer vision methods. Our proposal stands out by incorporating these features, showcasing significant advancements over existing solutions

Table 1. Short state-of-the-art survey.

Approach	Native API for smartphone	Native API for low-cost devices	Dataset Management	Multiple AI Facilities	Computer Vision
[Chan et al. 2013]	●	○	○	○	○
[Baldominos et al. 2014]	○	○	○	○	○
[Ribeiro et al. 2015]	○	○	○	●	○
[Li et al. 2017]	●	○	○	●	●
[Yao et al. 2022]	○	○	○	●	○
[De Caro et al. 2022]	○	○	○	●	●
[Shah et al. 2022]	○	○	○	●	○
[Lewicki et al. 2023]	○	○	○	○	○
[Fortuna et al. 2023]	○	○	○	●	○
[Guntupalli and Rudramalla 2023]	○	○	○	●	○
[Cerar and Hribar 2023]	○	○	○	●	○
[Lomonaco et al. 2023]	○	○	○	●	○
[Zhang et al. 2023]	○	○	○	●	●
[Hajipour et al. 2023]	○	○	○	●	○
[Merluzzi et al. 2023]	○	○	○	○	○
[Napisa et al. 2023]	○	○	○	●	○
[Baccour et al. 2023]	○	○	○	○	○
[Oliveira et al. 2024a]	○	○	○	●	○
[Nadar and Härri 2024]	○	○	○	●	○
Our Proposal	●	●	●	●	●

3. Proposal

In this section we present the rationale for the proposed AIaaS architecture considering its design and practical implementation. Also, we offer a comprehensive understanding of its functionalities, use-cases, and potential applications in real-world scenarios. The proposed architecture represents an innovative approach for providing AI resources as a service in edge computing. To achieve this, several technologies, contributions, and developments need to be systematically organized into research fronts. An overview of the proposed architecture is shown in Figure 2.

The hypothesis deduction is subdivided into three technological domains, from which activities are developed within the scope of the project.

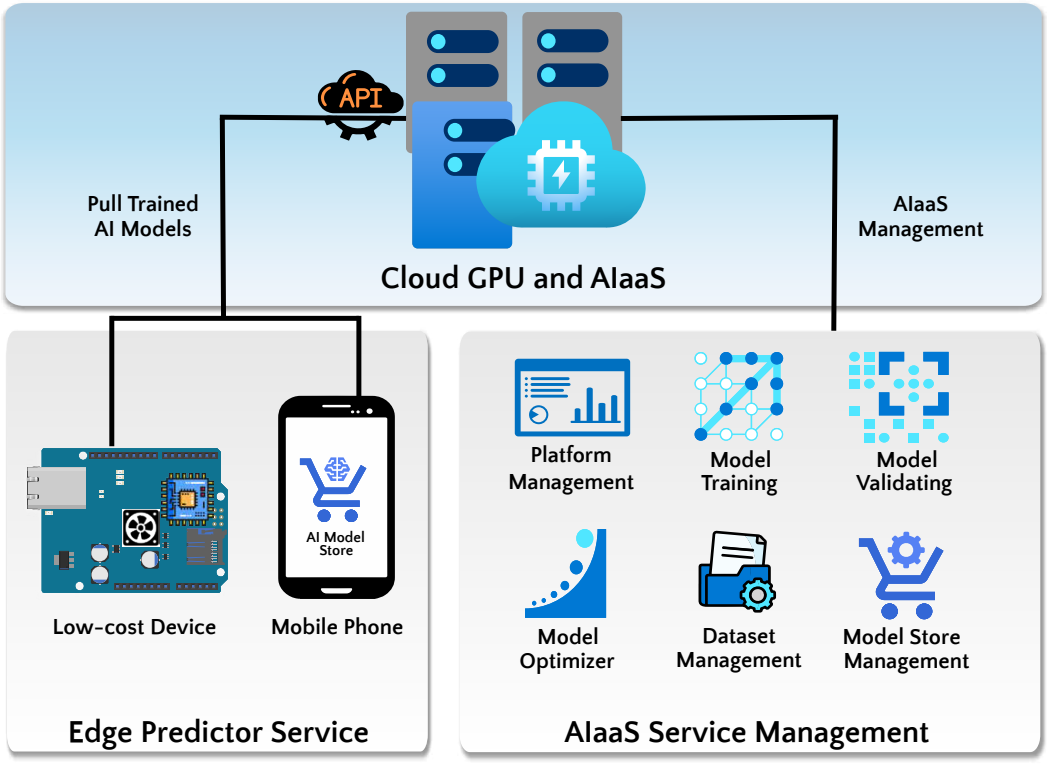

Figure 2. General Contribution: AIaaS Architecture.

1. **Infrastructure:** is based on cloud computing, and aims to provide services in the form of Infrastructure as a Service (IaaS) or Platform as a Service (PaaS). IaaS is a flexible and scalable cloud service that offers a complete computing infrastructure. PaaS is a comprehensive platform for application development and hosting, which accelerates the development process and reduces infrastructure costs. In this work, the AIaaS services of conventional providers available in the market were used.

2. **Service Management:** involves managing the essential functionalities of the proposed architecture. The API control and management mechanisms are developed and run on high-performance computing infrastructure.

3. **Edge Prediction:** encompasses the embodiment of all AI resources on demand on mobile devices or low-cost devices for prediction tasks. For example, medical image diagnosis can be supported using a mobile application to identify diseases and edge devices request the most suitable AI service from the infrastructure domain through an Application Programming Interface (API).

4. Results

This section summarizes the experimental evaluation of the proposed AIaaS architecture and demonstrates its effectiveness in delivering AI capabilities across diverse environments. The results, detailed in Chapters 5 and 6 of [Rodrigues Moreira 2024], demonstrate the potential of the architecture to enhance AI model deployment, performance, and accessibility while maintaining computational efficiency and scalability. These findings underscore the contribution of this thesis to the broader field of computing and its applicability to real-world scenarios.

Edge Predictor Service Evaluation. To assess the feasibility of deploying AI models on edge devices using the AIaaS Architecture, we conducted a series of experiments evaluating response time, energy consumption, and prediction accuracy. The results indicate that our architecture successfully integrates deep learning models into resource-constrained devices, enabling real-time inference, achieving a 35.7% reduction in inference latency compared to traditional cloud-based approaches. Furthermore, leveraging AIaaS for cognitive service retrieval enhances accessibility to advanced AI functionalities in remote and resource-limited areas [Rodrigues et al. 2023]. Moreover, AIaaS allows for seamless cognitive service acquisition from the AI Model Store, ensuring that low-cost edge devices benefit from advanced AI capabilities without requiring significant computational resources [Rodrigues Moreira et al. 2024c].

Service Management Evaluation. We evaluated the AIaaS Architecture in terms of its ability to manage the full lifecycle of AI models, including training, validation, and fine-tuning. We conducted performance benchmarks assessing the computational overhead imposed by AI training jobs on the resource pool, demonstrating that AIaaS efficiently allocates resources to optimize training performance while maintaining service availability. Additionally, a Federated Learning (FL) use case was implemented to validate the framework's capability to support decentralized AI training. The results indicated that AIaaS enables distributed model training with minimal performance degradation compared to centralized approaches, while also preserving data privacy, a critical aspect in today's data-sensitive applications [Rodrigues Moreira et al. 2024b].

Model Optimization and Adaptive Learning. To further enhance model performance, AIaaS incorporates optimization techniques such as grid search, random search, Bayesian optimization, and evolutionary strategies. We conducted a case study in medical image diagnosis, focusing on leukemia diagnosis using microscopy images. The results indicate that AIaaS-enabled optimization significantly enhances model accuracy, reducing misclassification rates by 14.6% compared to manually tuned models. These findings validate AIaaS's ability to refine AI models dynamically, adapting them to specific use cases and deployment environments [Rodrigues et al. 2022].

AIaaS Impact on Next-Generation Networks. AIaaS's integration into mobile network ecosystems, including Beyond 5G (B5G) and 6G, was analyzed to assess its role in enhancing network intelligence and adaptability. The proposed architecture aligns with 3rd Generation Partnership Project (3GPP)'s NWDAF, enabling seamless access to third-party cognitive services via a standardized northern interface. Experimental results show that AIaaS effectively supports network resource optimization, dynamic service orchestration, and real-time decision-making, contributing to improved Quality of Service (QoS) and enhanced network security. These findings demonstrate the architecture's potential to serve as a foundational component for AI-driven telecommunications infrastructures [Rodrigues Moreira et al. 2024a] [Oliveira et al. 2024b] [Rodrigues Moreira et al. 2025].

5. Contributions and Impact

The results of this work represent a significant advancement in AI deployment, particularly within the context of edge computing and next-generation network infrastructure. This study introduces a new AIaaS architecture that offers a comprehensive, scalable, and efficient solution for deploying AI cognitive services. The ability to optimize AI mod-

els and deploy them in resource-constrained environments, such as edge devices, is an important contribution to the AI community. Furthermore, this work has substantial implications for the development of intelligent applications in various industries, including healthcare, telecommunications, and beyond. The integration of AIaaS in mobile networks, such as 5G and 6G, positions this research as a key enabler for future network intelligence, enhancing AI's role of AI in next-generation telecommunications.

Overall, the findings of this thesis underscore the potential of AIaaS to support the management and deployment of AI models, thereby driving forward the efficiency and adaptability of AI systems across diverse applications. The proposed architecture paves the way for future research and development in this domain, offering a framework for future AI systems that are both scalable and capable of meeting the evolving demands of the industry, and paving the way for more efficient and adaptable solutions.

6. Research Accomplishments

This Thesis demonstrated productivity, as evidenced by the number and quality of the publications. Three high-impact journal articles were published: **Neurocomputing (IF 5.5)** [Rodrigues Moreira et al. 2025], **Applied Soft Computing (IF 7.2)** [Rodrigues Moreira et al. 2024a], and **Journal of Digital Imaging (IF 2.9)** [Rodrigues et al. 2022]. Additionally, the work was presented at prominent conferences, including two recent contributions at **IEEE CloudNet** and one at the **Workshop on 6G Networks (WG6)** coalocated event in the Brazilian Symposium on Computer Networks and Distributed Systems (**SBRC**). These papers have gained significant attention from the academic community, with a total of 62 citations.

This PhD research also involved significant collaborations with researchers from various institutions, including: Federal University of Viçosa (UFV), Federal University of Minas Gerais (UFMG), Federal University of Ceará (UFC), Federal University of Rio Grande do Sul (UFRGS), and University of Minho (UMinho) in Portugal. These collaborations resulted in five high-impact journal articles and 20 conference publications as co-author, resulting in 153 citations. Altogether, the contributions of this thesis led to 31 papers with a total of 215 citations[1].

7. Conclusion

The innovations in this thesis can be seen in how AIaaS architecture enables the easier and more efficient deployment of AI solutions, even in resource-limited settings. This reduces the complexity of managing AI infrastructure, allowing users to focus on utilizing AI capabilities without requiring technical knowledge. Additionally, AIaaS opens new opportunities for creating specialized AI applications, fostering advancements across various domains. Future work will aim to expand this architecture to support more diverse AI models and services by integrating cutting-edge technologies to further enhance its applicability and efficiency.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. André R. Backes gratefully

[1] All citations count were obtained from Google Scholar as of April 15, 2025.

acknowledges the financial support of CNPq (Grant #307100/2021-9). Bruno A. N. Travençolo is grateful to CNPq for the financial support (Grant #306436/2022-1). Larissa F. Rodrigues Moreira gratefully acknowledges the financial support of FAPEMIG (Grant #APQ00923-24).

References

[Ahmed et al. 2022] Ahmed, I., Jeon, G., and Piccialli, F. (2022). From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics*, 18(8):5031–5042.

[Baccour et al. 2023] 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.

[Baldominos et al. 2014] Baldominos, A., Albacete, E., Saez, Y., and Isasi, P. (2014). A scalable machine learning online service for big data real-time analysis. In *2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD)*, pages 1–8, Orlando, FL, USA. IEEE.

[Cerar and Hribar 2023] 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.

[Chan et al. 2013] Chan, S., Stone, T., Szeto, K. P., and Chan, K. H. (2013). PredictionIO: A Distributed Machine Learning Server for Practical Software Development. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2493–2496, New York, NY, USA. Association for Computing Machinery.

[De Caro et al. 2022] De Caro, V., Bano, S., Machumilane, A., Gotta, A., Cassarà, P., Carta, A., Semola, R., Sardianos, C., Chronis, C., Varlamis, I., Tserpes, K., Lomonaco, V., Gallicchio, C., and Bacciu, D. (2022). Ai-as-a-service toolkit for human-centered intelligence in autonomous driving. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 91–93, Pisa, Italy. IEEE.

[Fortuna et al. 2023] 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. In Shinkuma, R., Xhafa, F., and Nishio, T., editors, *Advances in Engineering and Information Science Toward Smart City and Beyond*, pages 53–73. Springer International Publishing, Cham.

[Guntupalli and Rudramalla 2023] 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 et al. 2023] 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.

[Jaffri and Sicular 2023] Jaffri, A. and Sicular, S. (2023). What's new in artificial intelligence from the 2023 gartner hype cycle.

[Lewicki et al. 2023] Lewicki, K., Lee, M. S. A., Cobbe, J., and Singh, J. (2023). Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

[Li et al. 2017] Li, L. E., Chen, E., Hermann, J., Zhang, P., and Wang, L. (2017). Scaling Machine Learning as a Service. In Hardgrove, C., Dorard, L., Thompson, K., and Douetteau, F., editors, *Proceedings of The 3rd International Conference on Predictive Applications and APIs*, volume 67 of *Proceedings of Machine Learning Research*, pages 14–29, Boston, USA. PMLR.

[Lomonaco et al. 2023] Lomonaco, V., Caro, V. D., Gallicchio, C., Carta, A., Sardianos, C., Varlamis, I., Tserpes, K., Coppola, M., Marmpena, M., Politi, S., Schoitsch, E., and Bacciu, D. (2023). AI-Toolkit: A Microservices Architecture for Low-Code Decentralized Machine Intelligence. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, Rhodes Island, Greece. IEEE.

[Luxton 2023] Luxton, R. (2023). Chapter 22 - Challenges and future aspects of sensor technology. In Barhoum, A. and Altintas, Z., editors, *Advanced Sensor Technology*, pages 853–877. Elsevier.

[Merluzzi et al. 2023] Merluzzi, M., Borsos, T., Rajatheva, N., Benczúr, A. A., Farhadi, H., Yassine, T., Müeck, M. D., Barmpounakis, S., Strinati, E. C., Dampahalage, D., Demestichas, P., Ducange, P., Filippou, M. C., Baltar, L. G., Haraldson, J., Karaçay, L., Korpi, D., Lamprousi, V., Marcelloni, F., Mohammadi, J., Rajapaksha, N., Renda, A., and Uusitalo, M. A. (2023). The Hexa-X Project Vision on Artificial Intelligence and Machine Learning-Driven Communication and Computation Co-Design for 6G. *IEEE Access*, 11:65620–65648.

[Nadar and Härri 2024] Nadar, A. and Härri, J. (2024). Enhancing Network Data Analytics Functions: Integrating AIaaS with ML Model Provisioning. In *2024 22nd Mediterranean Communication and Computer Networking Conference (MedComNet)*, pages 1–4, Nice, France. IEEE.

[Napisa et al. 2023] Napisa, K., Mababangloob, G. R., Lubag, M., Concepcion II, R., and Redillas, M. M. (2023). Explainable and Interpretable Artificial Intelligence as a Service for Green Smart Cities and Communities. In *2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–5, Coron, Palawan, Philippines. IEEE.

[Oliveira et al. 2024a] Oliveira, J., Almeida, J., Macedo, D., and Nogueira, J. (2024a). Um framework NWDAF para algoritmos de análise de dados de rede 5G e além. In *Anais do IV Workshop de Redes 6G*, pages 9–14, Porto Alegre, RS, Brasil. SBC.

[Oliveira et al. 2024b] Oliveira, J. M., Almeida, J., De Britto e Silva, E., Rodrigues Moreira, L. F., Moreira, R., Silva, F. O., Macedo, D. F., and Nogueira, J. M. (2024b).

Anomaly Detection Employing a 5G Core Data Analytics Framework. In *2024 IEEE 13th International Conference on Cloud Networking (CloudNet)*, pages 1–9.

[Ribeiro et al. 2015] Ribeiro, M., Grolinger, K., and Capretz, M. A. (2015). MLaaS: Machine Learning as a Service. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 896–902, Miami, FL, USA. IEEE.

[Rodrigues et al. 2022] Rodrigues, L. F., Backes, A. R., Travençolo, B. A. N., and de Oliveira, G. M. B. (2022). Optimizing a Deep Residual Neural Network with Genetic Algorithm for Acute Lymphoblastic Leukemia Classification. *Journal of Digital Imaging*, 35(3):623–637.

[Rodrigues et al. 2023] 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.

[Rodrigues Moreira 2024] Rodrigues Moreira, L. F. (2024). *Artificial Intelligence as a Service Architecture: an innovative approach for Computer Vision applications*. Phd thesis, Universidade Federal de Uberlândia, Uberlândia.

[Rodrigues Moreira et al. 2024a] Rodrigues Moreira, L. F., Moreira, R., de Oliveira Silva, F., and Backes, A. R. (2024a). Towards Cognitive Service Delivery on B5G through AIaaS Architecture. In *Anais do IV Workshop de Redes 6G*, pages 1–8, Porto Alegre, RS, Brasil. SBC.

[Rodrigues Moreira et al. 2024b] Rodrigues Moreira, L. F., Moreira, R., Martins, E. T., Jansen, V. F., Lima, Y. S., Rodrigues, L., Travençolo, B., and Backes, A. (2024b). Maximizing the power of cognitive services with an AI-as-a-Service architecture for seamless delivery. In *2024 IEEE 13th International Conference on Cloud Networking (CloudNet) (IEEE CloudNet 2024)*, page 8, Rio de Janeiro, Brazil. IEEE.

[Rodrigues Moreira et al. 2025] Rodrigues Moreira, L. F., Moreira, R., Travençolo, B. A. N., and Backes, A. R. (2025). Deep learning based image classification for embedded devices: A systematic review. *Neurocomputing*, 623:129402.

[Rodrigues Moreira et al. 2024c] Rodrigues Moreira, L. F., Saar, L., Moreira, R., Rodrigues, L., Travençolo, B., and Backes, A. (2024c). Enabling intelligence on edge through an artificial intelligence as a service architecture. In *2024 IEEE 13th International Conference on Cloud Networking (CloudNet) (IEEE CloudNet 2024)*, page 8, Rio de Janeiro, Brazil. IEEE.

[Shah et al. 2022] 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.

[Yao et al. 2022] Yao, T., Wang, J., Wan, M., Xin, Z., Wang, Y., Cao, R., Li, S., and Chi, X. (2022). VenusAI: An artificial intelligence platform for scientific discovery on supercomputers. *Journal of Systems Architecture*, 128:102550.

[Zhang et al. 2023] 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.