

Multi-Partner Project: dAIEDGE - A Network of Excellence for Distributed, Trustworthy, Efficient and Scalable AI at the Edge

Alain Pagani^{*}, José Cano[†], Haralampos-G. Stratigopoulos[‡], Aysajan Abidin[§], Mhd Rashed Al Koutayni^{*}, Luca Benini[¶], Angelos Bilas^{||}, Alessandro Capotondi^{**}, Roberto Cavicchioli^{**}, Brian Clerkin^{††}, Oscar Deniz^{‡‡}, Margaux Divernois^x, Baptiste Dupertuis^x, Dorvan Favre^x, Giulio Gambardella^{††}, Ander Garcia Gangoit^x, Carlo Augusto Grazia^{**}, Dominik Günzel^{xii}, Jude Haris[†], Klodjan K. Hidri^{||}, Maïck Huguenin^x, Manal Jammal^{xiii}, Paul Kling[‡], Christos Kozanitis^{||}, Xavier Lessage^{xiv}, Srikanth Mandapati^{xii}, Philippe Massonet^{xiv}, Alfio Di Mauro[¶], Varesh Mishra[§], Juan Odriozola^{xi}, Javier Parra Domínguez^{xiii}, Nuria Pazos^x, Viviane Potocnik[¶], Miguel de Prado^{xv}, Rohit Prasad^{xvi}, Spyridon Raptis[‡], Gregoire Rebstein^x, Ignacio Sañudo Olmedo^{xviii}, Mohamed Selim^{*}, Chinmay Satish Shrivastav^{**}, Noelia Vallez^{‡‡}, Giorgos Vasiliadis^{|| xvii}, Micaela Verrucchi^{xviii}, Enrico Vincenzi^{xviii}, Damian Vizár^{xix}, Devendra Vyas^{xv}, Stefan Wiehle^{xii}

^{*}DFKI, Germany [†]University of Glasgow, UK [‡]Sorbonne Université, CNRS, LIP6, France [§]KU Leuven, Belgium
[¶]ETH Zürich, Switzerland ^{||}FORTH, Greece ^{**}University of Modena and Reggio Emilia, Italy ^{††}Synopsys, Ireland
^{‡‡}University of Castilla-La Mancha, Spain ^xHES-SO, Switzerland ^{xi}Vicomtech, Spain ^{xii}DLR, Germany
^{xiii}University of Salamanca, Spain ^{xiv}CETIC, Belgium ^{xv}VERSES, Switzerland ^{xvi}Université Paris-Saclay, CEA, List, France
^{xviii}Hellenic Mediterranean University, Greece ^{xviii}Hipert s.r.l., Italy ^{xix}CSEM, Switzerland

Abstract—The dAIEDGE Network of Excellence (NoE) seeks to strengthen and support the development of a dynamic European cutting-edge Artificial intelligence (AI) ecosystem under the umbrella of the European Lighthouse for AI, and to sustain the development of advanced AI. dAIEDGE fosters the exchange of ideas, concepts, and trends on cutting-edge next generation AI, creating links between ecosystem actors to help both the European Commission (EC) and the European Union (EU) and the peripheral AI constituency identify strategies for future developments in Europe. Our main objective is to advance Europe’s innovation and technology base by developing a comprehensive policy and governance approach to AI in order for the EU to become a world leader in innovation in the data economy and its applications.

I. INTRODUCTION

Artificial intelligence (AI) is moving from centralised data centres to the distributed environments where data is produced. This shift towards computing at the edge enables faster and more energy-efficient decision making and supports privacy preservation by reducing the transfer of sensitive data. Yet, the deployment of AI directly on embedded devices remains challenging [1]. Real-time operation, limited computational and energy resources, and the need for explainable and verifiable behaviour require a new generation of design methodologies that

This work was funded by the European Network of Excellence dAIEDGE under Grant Agreement N^o 101120726.

integrate algorithms, hardware, and system orchestration across the entire edge-to-cloud continuum. Developing such systems demands not only technical innovation but also collaboration between diverse disciplines, from microelectronics and computer vision to distributed computing and human-centred AI.

The dAIEDGE project was initiated to tackle these challenges and strengthen Europe’s leadership in edge AI technologies. Section II outlines the project’s structure, objectives, and vision, while the subsequent sections discuss key concepts, results, ongoing work, and lessons learned throughout the project.

II. THE dAIEDGE PROJECT

dAIEDGE is a Horizon Europe project funded under the AI, Data, and Robotics Partnership and coordinated by the German Research Center for Artificial Intelligence (DFKI). Running from September 2023 to August 2026, it brings together thirty-six partners from twelve European countries, including universities, research organisations, technology providers, and innovation hubs. Its overarching vision is to create a sustainable and competitive Network of Excellence (NoE) that connects scientific research, industrial development, and policy initiatives in developing hardware-aware, energy-efficient, and trustworthy AI for deployment in real-world applications.

The project’s work is guided by a coherent set of objectives that define a comprehensive strategy for advancing edge AI in Europe. At its foundation, the project builds a dynamic ecosystem that mobilises stakeholders and promotes collaboration among researchers, innovators, and policymakers. Through this network, it creates a platform for knowledge exchange, exploration of new paradigms, and coordinated development of edge AI technologies and standards. The project reinforces this ecosystem through the development of a Strategic Research Agenda (SRA) for Edge AI in Europe, to be published in early 2026. This agenda synthesises research trends, identifies technological and methodological gaps, and defines priorities for future innovation. It integrates insights from disciplines such as cognitive and neuromorphic AI, swarm intelligence, and distributed learning, while addressing cross-cutting challenges including energy efficiency, contextual computing, virtualisation, and the edge-to-cloud continuum. The SRA is being developed in close collaboration with the wider European AI community, ensuring that its recommendations extend beyond the dAIEDGE consortium and contribute to shaping Europe’s future strategy for edge AI.

To demonstrate tangible technological value, the project applies a lead-by-demonstration approach in which its scientific developments and theoretical foundations are practically validated in representative industrial domains, such as mobility, robotics, and smart manufacturing. These demonstrators focus on privacy, safety, security, and sustainability and show how distributed AI can support a digitally driven, climate-neutral European economy. The demonstration strategy is complemented by open calls that allow additional partners to join the network and extend its technical and application scope.

The project further enhances collaboration across Europe by aligning its activities with other partnerships, NoEs, and national initiatives. It coordinates the development of open AI platforms and frameworks, ensuring that research and innovation outcomes are interoperable and reusable, and supporting Europe’s autonomy in key edge technologies. A key element of this collaboration is the dAIEDGE Virtual Laboratory (dAIEDGE-VLab), a distributed online environment for remote AI benchmarking and experimentation on diverse edge devices, described in more detail in Section V.

Beyond its technical achievements, the project invests in community building, dissemination, skills development, education and training. It raises awareness of the technological and societal importance of edge AI and ensures that the outcomes are widely accessible. The communication and outreach activities create a long-term structure for collaboration that will persist beyond the project’s duration and help shape future European strategies for AI and digital transformation.

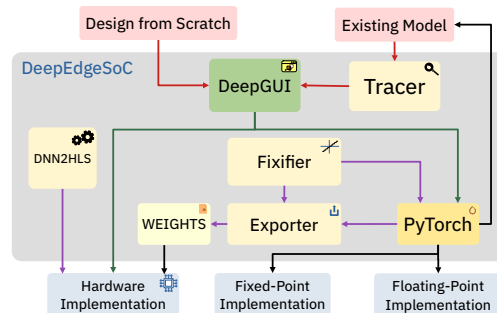


Figure 1: DeepEdgeSoC framework diagram.

Through these integrated activities, the project delivers scientific advances and practical enablers that strengthen Europe’s capability in edge AI. It provides open-source tools and open reference architectures, benchmark datasets to enable comparisons and stimulate joint research, toolchains for hardware-software co-design creating a unified design flow, orchestration mechanisms for dynamic workload balancing across edge-cloud networks, and methodologies for verification, reliability, safety, security, and trust throughout the AI lifecycle. The project further explores model compression, quantisation, and adaptive inference to ensure efficient execution under strict latency and power constraints. Together, these outcomes lay the foundation for autonomous, efficient AI at the edge. For the embedded systems and design automation communities, the project offers an opportunity to evaluate hardware-aware AI implementations on heterogeneous platforms including CPUs, GPUs, FPGAs, and emerging RISC-V and neuromorphic architectures, defining the next generation of dependable and energy-efficient AI systems.

III. AI HARDWARE ACCELERATORS

Efficient hardware accelerators are vital for deploying AI workloads on edge devices with tight power and latency constraints. We explored approaches from FPGA frameworks to reconfigurable arrays and hardware-software co-design toolkits, each addressing different aspects of performance, flexibility, and automation.

A. DeepEdgeSoC

DeepEdgeSoC [2] (Fig. 1) is an FPGA-based framework for deploying deep neural networks (DNNs) on System-on-Chip (SoC) platforms for energy- and latency-constrained edge AI workloads. It integrates model design, quantisation, compression, and hardware prototyping. Validated on multiple vision tasks, DeepEdgeSoC achieves a $\sim 10\times$ reduction in model size and a $\sim 22\times$ improvement in energy efficiency on an AMD Xilinx ZCU102 FPGA SoC [3]. The face detection network was extended to estimate landmarks, head pose, and driver actions, forming a driver monitoring system.

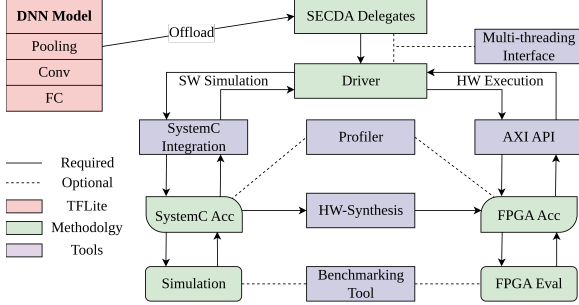


Figure 2: Overview of the SECDA-TFLite toolkit [4].

B. SECDA-TFLite and SECDA-LLM

SECDA-TFLite [4] (Fig. 2) and SECDA-LLM [5] are hardware-software co-design toolkits for rapid development and evaluation of FPGA-based AI accelerators for the TensorFlow Lite and llama.cpp frameworks, respectively. They enable iterative simulation-based design following the SECDA design methodology [6]. Both toolkits include a benchmarking tool for rapid model testing, an automated hardware generation tool using HLS/HLX for FPGA boards, remote server support, VSCode DevContainers installation, a profiler, and delegate/backend generation for 11 DNN/LLM operations.

C. NX-CGRA

NX-CGRA [7] is a software-programmable Coarse-Grained Reconfigurable Array for accelerating diverse AI workloads at the edge. Its 2D mesh of heterogeneous processing elements (PEs), connected via a switchless torus interconnect, supports a MIMD execution model for irregular computation graphs in transformers [8], convolutional neural networks (CNNs), and lidar-based detection. PEs perform INT8-FP32 arithmetic with vector multiply-accumulate (MAC) units, while memory operation blocks (MOBs) manage address computation and data movement to reduce latency. Integrated into an open-source SoC [9], NX-CGRA demonstrates strong performance in CNNs, transformers, and lidar fusion.

IV. AI HARDWARE RELIABILITY

Hardware-level reliability is critical for AI chips because they operate under extreme performance, density, and energy constraints, and even small hardware faults can lead to major degradation in accuracy, safety, or system trustworthiness. In this context, we conducted error sensitivity analysis on state-of-the-art text embedding transformer models targeting different applications [10]. Interestingly, the error injection campaign identifies that mainly a portion of the self-attention layer, specifically the Value tensor, is very sensitive to single bit-flips. These results pave the way for a more effective error tolerance approach in future transformer-based embeddings and possibly other self-attention based models.

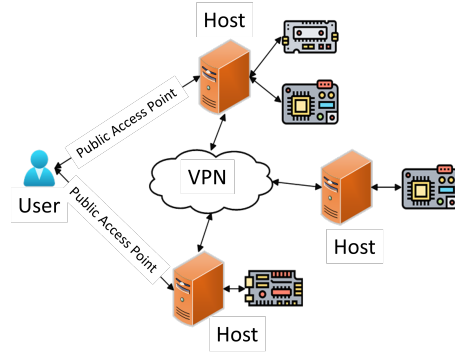


Figure 3: dAIEdge-VLab.

V. DAIEDGE-VLAB

Benchmarking AI models is essential for responsible AI development, providing standardised metrics like accuracy, speed, and efficiency. Benchmarking across diverse edge platforms is challenging due to hardware and software variability. The dAIEdge-VLab [11]–[14] (Fig. 3) tackles these issues by offering a distributed virtual laboratory for remote AI benchmarking and experimentation on various edge devices. It democratises embedded AI development by allowing users to run live experiments, such as benchmarking, on-device training, and federated learning (FL), without requiring deep hardware expertise. To our knowledge, no vendor-neutral, multi-platform, scalable online benchmarking system currently exists, motivating dAIEdge-VLab’s creation.

A first version of the dAIEdge-VLab is currently operational (<https://vlab.daiedge.eu/>), supporting multiple devices such as Raspberry Pi, NVIDIA Jetson boards, STM32-based MCUs, RISC-V-based MCUs, NPUs, SNN SoCs, and FPGA-based accelerators, as well as different ML toolchains like TensorFlow Lite (-micro), ONNX Runtime, Cube.AI, TensorRT, and OpenVino.

VI. AI SECURITY

A. Enclave-Aware Model Partitioning

As AI inference moves to untrusted edge environments, both input data and model parameters must remain confidential. Trusted Execution Environments (TEEs) like Intel SGX provide strong hardware isolation but are constrained by limited secure memory (typically < 128 MB) and expensive transitions between trusted and untrusted spaces. To address these limitations, we propose a privacy-preserving inference framework that runs entirely within an SGX enclave [15]. The system establishes a secure, attested channel between the client and enclave, and introduces an enclave-aware model partitioning mechanism that decomposes DNNs into independently executable subgraphs fitting within the enclave’s limited memory. Each partition is securely loaded, executed, and released in sequence, while intermediate activations remain confined within the enclave.

boundary. This approach allows efficient use of limited enclave memory, reducing paging overhead and maintaining data locality, achieving up to 4× performance gains over naïve full-model execution.

B. Secure Federated Learning at the Edge

FL enables multiple clients to collaboratively train a model by sharing only updates with a central server, ensuring data remains local. However, sharing model updates can still expose sensitive information or enable backdoor insertion, allowing adversaries to reconstruct private data, infer membership, or recover local models of participants. We proposed a secure Edge AI-optimised FL architecture [16] for the edge using the Flower framework [17], integrating security countermeasures, such as anomaly detection, differential privacy, secure Multi-Party Computation (MPC), homomorphic encryption, and blockchain. Each mitigates specific risks, and their combination provides strong layered defence, reducing adversarial success from over 70% to below 10–15%.

C. Encrypted Federated Learning

MPC-based FL frameworks mitigate model inference risks by distributing local updates among multiple aggregators but face challenges such as synchronisation needs and high communication overhead. In [18], we present the first evaluation of ASCON, NIST’s new lightweight cryptography standard, in MPC. Our approach encrypts local models with ASCON, stores them externally, and splits encryption keys into secret shares encrypted for each aggregator, which then collaboratively aggregate and encrypt the global model.

D. Integrating MPC and TEE

MPC offers strong privacy but high computational cost, especially for nonlinear functions, while TEE provides efficiency but relies on hardware trust and is prone to side-channel attacks. We propose a hybrid MPC-TEE workflow combining their strengths to enhance privacy, efficiency, and security. The neural network is split into linear and nonlinear layers: linear operations are computed collaboratively via MPC, while a randomly chosen party performs nonlinear computations inside a TEE. The output is then re-shared using MPC, repeating this process for each layer.

E. Secure model update over the air

Secure model updates over the air (MUOTA) are essential for maintaining edge AI security and reliability throughout deployment. However, ensuring long-term confidentiality is challenging due to physical and software attacks (e.g., debug port resurrection, side-channel leaks, BLE exploits) that may expose the decryption key, a high-value asset compromising all future model updates and also enabling adversarial query crafting.

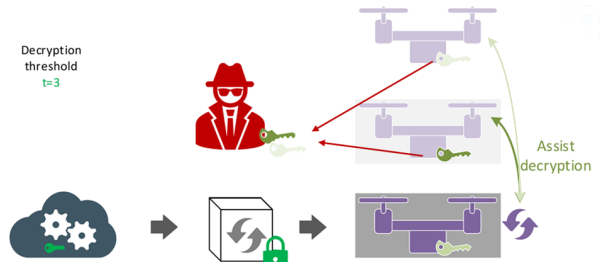


Figure 4: Secure MUOTA with drone as Edge device.

To address this, we propose a confidentiality-hardened MUOTA architecture using threshold symmetric encryption (Fig. 4), extending the baseline in [19]. Each edge device holds only a key share and must interact with $t - 1$ neighbouring devices via short-range communication protocol to decrypt incoming models, where t is a configurable threshold. A higher t increases the number of compromised devices required for key recovery, and when coupled with intrusion detection, effectively limits the likelihood of such event.

VII. SUSTAINABLE AI

dAIEDGE advances sustainable AI through governance and life cycle assessment approaches that evaluate environmental impacts from training to disposal, beyond mere energy use. Guided by Green AI principles [20], it promotes energy transparency, efficiency over brute-force accuracy, and standardised carbon accountability [21]. Edge AI sustainability emphasises distributed, energy-efficient systems that process data locally on resource-constrained devices, reducing emissions from data transfer. This aligns with Europe’s digital and green transitions by enabling energy-conscious, socially responsible AI. dAIEDGE also calls for broader sociotechnical metrics consistent with Triple Bottom Line sustainability, integrating environmental, social, and economic dimensions, such as digital inclusion and fair labour, into evaluation frameworks. This holistic perspective supports regulatory innovation and sustainability disclosures to institutionalise greener AI [22], [23].

VIII. RESOURCE ALLOCATION

Edge AI infrastructures face resource allocation challenges similar to those in datacentre and cloud environments [24], [25]. In heterogeneous testbeds, such as the dAIEdge-VLab, diverse devices with varying performance and energy profiles complicate manual resource selection. To address this, we introduce k3sRalloc, a Kubernetes Operator for Lightweight Kubernetes (K3s) clusters that automates cost-efficient, latency-aware deployment of ML inference workloads. Users specify only the model location and target latency, and k3sRalloc selects the least costly node meeting the constraint. It extends the K3s control plane with a custom resource

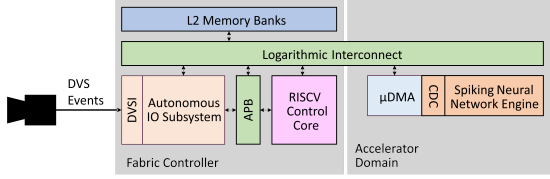


Figure 5: Simplified event-driven interface for SNNs.

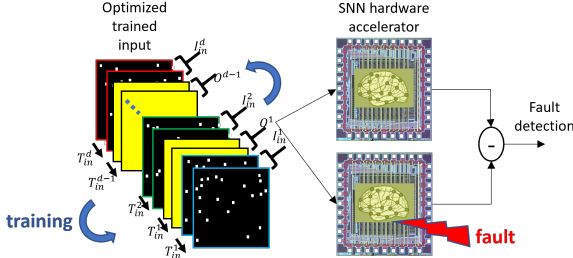


Figure 6: Training input to detect faults in SNN hardware.

definition (CRD), maintains latency profiles across devices, benchmarks unprofiled models, and ranks nodes by cost–performance trade-offs. Over time, k3sRalloc refines its data using predictive techniques such as matrix completion to improve accuracy and scalability.

IX. SPIKING NEURAL NETWORKS

A. Event-Driven Sensor Interfaces

We develop a hardware/software interface framework that bridges dynamic-vision (DVS) and frame-based sensors with neuromorphic processors, enabling asynchronous, data-dependent, energy-proportional sensing and computation at the edge. As illustrated in Fig. 5, the interface converts DVS streams into a memory-mapped representation through a μ DMA engine, bypassing CPU control and minimising wake-ups. The interface is fully supported by the open-source PULP-SDK and integrated with PyTorch-based training pipelines, enabling quantisation-aware deployment of spiking models at the edge. It was validated in silicon within the heterogeneous RISC-V Kraken SoC [26]. The results demonstrate high-throughput SNN workloads while keeping power consumption tightly coupled to sensor activity.

B. Testing of SNN hardware accelerators

With neuromorphic processors entering early commercialisation, we address the challenge of testing these specialised chips [27]. Fault injection shows that faults in critical synapses or neurons can significantly degrade accuracy [28], [29]. We propose a test generation algorithm that trains a spiking-domain test input to detect faults, replacing time-consuming fault simulations with objective functions that maximise fault sensitisation and output propagation. Compared to existing methods, our approach achieves perfect critical fault coverage with shorter test duration and generation time. The compact

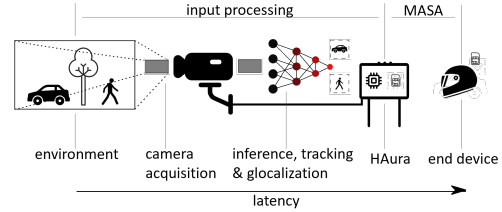


Figure 7: HAura edge-to-end architecture within MASA.

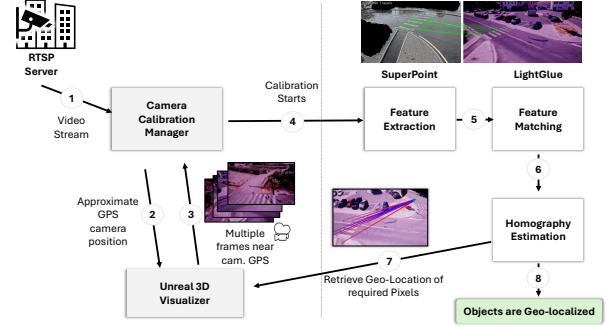


Figure 8: Object geolocation in traffic camera imagery.

test can be stored on-chip for efficient in-field testing and is validated on Intel’s Loihi 2 processor.

C. Security for SNNs

We address two threat models: adversarial attacks and hardware Trojans. Adversarial attacks introduce subtle input perturbations to mislead SNNs. We proposed two attack algorithms [30]: an input-specific attack crafting per-sample perturbations and a universal attack effective across inputs. Experiments on standard neuromorphic datasets show superior attack success rate, stealthiness, and generation time over state-of-the-art methods. The proposed hardware Trojan [31] embeds a “Trojan neuron” triggered by a crafted spiking input, causing permanent saturation and network corruption. Implemented on a digital SNN accelerator, it shows negligible area and power overhead, enabling stealthy insertion.

X. SMART CITY APPLICATIONS

A. Awareness for Vulnerable Road Users

We aim to showcase a real-time cooperative perception system that enhances urban safety by detecting Vulnerable Road Users (VRUs) in areas outside a driver’s or vehicle’s line of sight. The system links the HAura road-side infrastructure developed by Hipert, performing distributed AI-based detection and tracking, with the AEGIS Rider smart helmet, which provides augmented reality (AR) alerts to a motorbike rider. The use case is integrated within the Modena Automotive Smart Area (MASA) [32]. The goal is to prove that edge-to-device communication can achieve low-latency situational awareness under 150 ms, enabling real-time reaction to potential hazards in blind spots.

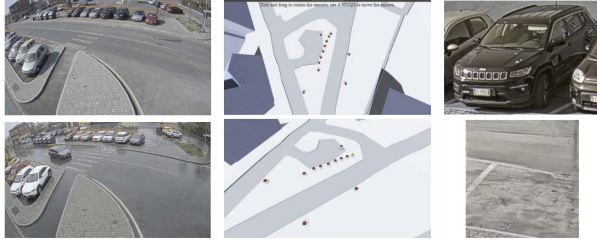


Figure 9: Parking space visualisation and detection.

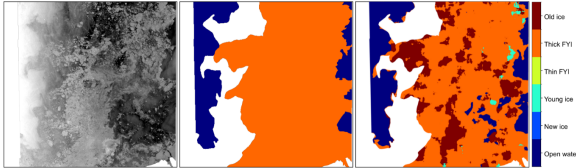


Figure 10: Sea ice classification on a test scene in the AI4Arctic dataset; left: original, middle: reference, right: result.

B. Parking Lot Management

This work presents an automatic, scalable parking lot management pipeline using smart city cameras, designed to operate without external sensors or manual calibration, and demonstrated in MASA. The system runs directly on next-generation urban cameras (e.g., Hipert’s HAura), enabling fully autonomous deployment in real-world traffic scenarios. The framework comprises two main components: (1) an end-to-end system that automates camera and object geolocation without explicit calibration (see Fig. 8), and (2) a Parking Lot Management (PLM) application for parking detection and occupancy classification (see Fig. 9). Together, these modules provide a fully automated, scalable solution for real-time parking monitoring.

C. Automatic firearm detection

Urban gun violence remains a major concern worldwide, with increasing incidents involving lone shooters. Our work focuses on bridging the gap between laboratory research and large-scale, real-time deployment through edge-based visible and concealed weapon detection solutions. For concealed weapon, we selected a miniature thermal camera that can be connected to an Android smartphone (targeted for a chest-worn camera for a police officer). A Yolo-based detector was trained on a thermal dataset. For visible RGB cameras we use both Yolo-based and pose-recognition. More details are available in [33], [34].

XI. SPACE APPLICATIONS

Satellite Earth Observation (EO) provides vital information on earth processes; here, we focus on monitoring drifting sea ice with Synthetic Aperture Radar (SAR). Currently, data transfer delays limit timely access, which onboard processing can overcome. We implemented an onboard processing chain on an AMD Zynq UltraScale+

ZU19EG MPSoC [35], including SAR image formation using concurrently implemented Omega-K and chirp scaling algorithms, followed by a VGG13-based CNN for sea-ice classification. With 70% pruning and 8-bit quantisation, the model achieved an F1 score of 80.6%, 38 s latency, and 21W power consumption. A classification example is shown in Fig. 10.

We also proposed an innovative Machine Learning Operations (MLOps) platform [36], designed to address the limited computational resources constraints at the edge, and validated on a satellite-based sea ice classification test bed. The platform continuously enhances model accuracy through server labelling and retraining processes, automating updates while ensuring stringent deployment quality standards. It integrates modern tools, such as Apache Airflow [37] for orchestration, MLflow [38] for experiment tracking and model management, and Apache TVM [39] for model optimisation, into a unified framework, managing the lifecycle of AI models.

XII. AGENTIC EDGE AI FOR IMMERSIVE AND COLLABORATIVE SMART WAREHOUSE

Warehouse automation is moving beyond task-specific robotics and towards agentic AI systems that perceive, reason, and collaborate with humans, facilitating adaptive, human-centred warehouse ecosystems. We address the central research question on how to embed autonomy directly at the edge, making robots act as adaptive agents on three representative scenarios: a) monitoring with smart cameras and drones, b) adaptive navigation of autonomous mobile robots, and c) human-robot collaboration through gestures and shared tasks.

Our methodology integrates three innovation layers. A perception pipeline ensures robust context awareness. Agentic reasoning and planning modules support exploration, routing, and adaptive decision-making. Human-centred interfaces, including XR/AR helmets for contextual guidance and body tracking, enable seamless collaboration. Deployed across the compute continuum, from deep edge perception to meta-edge coordination, the system demonstrates resilient smart agents operating across physical and digital environments.

XIII. CONCLUSION

AI is shifting from centralised data centres to distributed edge environments for faster, greener, and more secure decision-making. Yet, edge AI deployment remains complex, requiring integrated design across algorithms, hardware, and systems. The dAIEDGE project tackles these challenges, advancing Europe’s edge AI leadership. In this paper, we presented its vision and objectives, highlighting achievements, ongoing developments, and insights gained throughout its development.

REFERENCES

- [1] P. Gibson, J. Cano, E. Crowley, A. Storkey, and M. O'boyle, "DLAS: A Conceptual Model for Across-Stack Deep Learning Acceleration," *ACM Transactions on Architecture and Code Optimization*, vol. 22, no. 1, 2025.
- [2] M. R. Al Koutayni, G. Reis, and D. Stricker, "DeepEdgeSoC: End-to-end deep learning framework for edge IoT devices," *Internet of Things*, vol. 21, 2023.
- [3] —, "Optimization strategies for neural network deployment on FPGA: An energy-efficient real-time face detection use case," *Internet of Things*, vol. 33, 2025.
- [4] J. Haris, P. Gibson, J. Cano, N. B. Agostini, and D. Kaeli, "SECD-A-TFLite: A toolkit for efficient development of FPGA-based DNN accelerators for edge inference," *Journal of Parallel and Distributed Computing*, vol. 173, pp. 140–151, 2023.
- [5] J. Haris, R. Saha, W. Hu, and J. Cano, "Designing Efficient LLM Accelerators for Edge Devices," *arXiv:2408.00462*, 2024.
- [6] J. Haris, P. Gibson, J. Cano, N. B. Agostini, and D. Kaeli, "SECD-A: Efficient hardware/software co-design of FPGA-based DNN accelerators for edge inference," in *Proc. IEEE 33rd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, 2021, pp. 33–43.
- [7] R. Prasad, "Coarse-grained reconfigurable array architecture and associated method," Jun. 2025. [Online]. Available: <https://patents.google.com/patent/EP4564187A1>
- [8] —, "An ultra-low-power cgra for accelerating transformers at the edge," Jan. 2025, working paper or preprint. [Online]. Available: <https://hal.science/hal-04914400>
- [9] S. Machetti, P. D. Schiavone, T. C. Müller, M. Peón-Quirós, and D. Aienza, "X-HEEP: An open-source, configurable and extendible RISC-V microcontroller for the exploration of ultra-low-power edge accelerators," *arXiv:2401.05548*, 2024.
- [10] S. Sapkal, U. Zahid, G. Gambardella, H.-G. Stratigopoulos, and B. Clerkin, "On the fault sensitivity of natural language embeddings computation," in *Proc. IEEE European Test Symposium (ETS)*, 2025.
- [11] M. Huguenin, B. Dupertuis, R. Frund, M. Divernois, and N. Pazos, "Online AI benchmarking on remote board farms," in *Proc. European Conference on Edge AI Technologies and Applications (EEAI)*, 2024.
- [12] B. Dupertuis *et al.*, "Scalable virtual lab for remote experiments on AI-powered edge devices," in *HiPEAC*, 2025.
- [13] R. López Blanco *et al.*, "Enhancing AI benchmarking in dAIEdge-VLab with blockchain," in *Proc. International Conference on Distributed Computing and Artificial Intelligence (DCAI)*, 2025.
- [14] B. Dupertuis *et al.*, "Fair AI experimentation on edge device clusters via distributed orchestration in dAIEdge-VLab," in *Proc. European Conference on Edge AI Technologies and Applications (EEAI)*, 2025.
- [15] K. Papafragkaki and G. Vasiliadis, "InferONNX: Practical and privacy-preserving machine learning inference using trusted execution environments," in *Detection of Intrusions and Malware, and Vulnerability Assessment*, M. Egele, V. Moonsamy, D. Gruss, and M. Carminati, Eds. Cham: Springer Nature Switzerland, 2025, pp. 25–43.
- [16] X. Lessage *et al.*, "Secure federated learning applied to medical imaging with fully homomorphic encryption," in *Proc. IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, 2024.
- [17] "Flower: A friendly federated learning framework." [Online]. Available: <https://flower.ai>
- [18] P. Schwarz, E. Pohle, A. Abidin, and B. Preneel, "Evaluating Ascon in secure multi-party computation using reverse multiplication-friendly embeddings," in *Proc. ACM Workshop on Privacy in the Electronic Society (WPES)*, 2025.
- [19] N. Deligiannakis *et al.*, "D4.2 – Cognitive cloud software-ized infrastructure customization (connectors). HYPER-AI project deliverable, Grant Agreement No. 101135982." <https://hyper-ai-project.eu/>, 2025.
- [20] R. Schwartz *et al.*, "Towards a standard for identifying and managing bias in artificial intelligence," *Special Publication (NIST SP)*, National Institute of Standards and Technology, Gaithersburg, MD, 2022.
- [21] J. Dodge *et al.*, "Measuring the carbon intensity of AI in cloud instances," in *Proc. ACM Conference on Fairness, Accountability, and Transparency*, 2022.
- [22] K. Chan *et al.*, "Greening AI: a policy agenda for the artificial intelligence and energy revolutions," https://eprints.lse.ac.uk/123705/1/Chan_greeningaipublished.pdf, 2024.
- [23] O. O. Oyewole and J. F. Joseph, "Sustainable AI and green computing: Reducing the environmental impact of large scale models with energy efficient techniques," *International Journal of Scientific Research in Network Security and Communication*, vol. 13, no. 3, 2025.
- [24] Y. Sfakianakis, C. Kozanitis, C. Kozyrakis, and A. Bilas, "Quman: Profile-based improvement of cluster utilization," *ACM Transactions on Architecture and Code Optimization*, vol. 15, no. 3, 2018.
- [25] C. Delimitrou and C. Kozyrakis, "Paragon: QoS-aware scheduling for heterogeneous datacenters," in *Proc. 11th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2013, p. 77–88.
- [26] V. Potocnik *et al.*, "Kraken: An open-source RISC-V SoC for ultra-low power multi-modal perception," *Research Square*, 2024.
- [27] S. Raptis and H.-G. Stratigopoulos, "Minimum time maximum fault coverage testing of spiking neural networks," in *Proc. Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2025.
- [28] T. Spyrou, S. Hamdioui, and H.-G. Stratigopoulos, "SpikeFI: A fault injection framework for spiking neural networks," *arXiv:2412.06795*, 2024.
- [29] Z. Jouni and H.-G. Stratigopoulos, "STDP-trained spiking neural network reliability assessment through fault injections," in *Proc. IEEE International Symposium on On-Line Testing and Robust Systems (IOLTS)*, 2025.
- [30] S. Raptis and H. G. Stratigopoulos, "Input-specific and universal adversarial attack generation for spiking neural networks in the spiking domain," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2025.
- [31] S. Raptis, P. Kling, I. Kaskampas, I. Alouani, and H.-G. Stratigopoulos, "Input-triggered hardware trojan attack on spiking neural networks," in *Proc. IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2025.
- [32] G. Ferraro *et al.*, "Towards Smart Cities with AI-CAM: Assisted by Infrastructure Cooperative Awareness Messages," in *Proc. IEEE Conference on Standards for Communications and Networking (CSCN)*, 2025.
- [33] J. Muñoz, H. Albandea, J. Ruiz-Santaquiteria, O. Deniz, and M. Verucchi, "Experiences in deploying a weapon detector in a smart city," in *Proc. European Conference on Edge AI Technologies and Applications (EEAI)*, 2025.
- [34] J. D. Muñoz, J. Ruiz-Santaquiteria, O. Deniz, and G. Bueno, "Concealed weapon detection using thermal cameras," *Journal of Imaging*, vol. 11, no. 3, 2025.
- [35] D. Günzel and S. Mandapati, "Satellite on-board processing of synthetic aperture radar data for rapid delivery of latency sensitive maritime information products," in *Active and Passive Remote Sensing of Oceans, Seas, and Lakes*, vol. 13264, 2025.
- [36] J. Odriozola, G. Paolini, M. Flores, and A. Garcia, "MLOps for edge AI: Satellite sea ice detection test bed," in *Proc. 22nd International Conference on Distributed Computing and Artificial Intelligence (DCAI)*, 2025.
- [37] "Apache airflow: A platform to programmatically author, schedule and monitor workflows." [Online]. Available: <https://airflow.apache.org>
- [38] "MLFlow: A developer platform to build AI applications and models with confidence." [Online]. Available: <https://mlflow.org>
- [39] "Apache TVM: An open machine learning compiler framework." [Online]. Available: <https://tvm.apache.org>