



D5.1 – Algorithms and methods prototype



Project Title	A network of excellence for distributed, trustworthy, efficient and scalable AI at the Edge
Project Acronym	dAIEDGE
Grant Agreement No	101120726
Topic	HORIZON-CL4-2022-HUMAN-02-02
Start Date of Project	September 1 st , 2023
Duration of Project	36 Months

Name of the Deliverable	Algorithms and Methods prototype
Number of the Deliverable	D5.1
Related WP Number and Name	WP5 - Edge AI Technological Advances for Cross-fertilisation
Related Task Number and Name	T5.1 - Edge AI algorithms and methods for resource-constrained devices
Deliverable Dissemination Level	PU - Public
Deliverable Due Date	M16 – December 2024
Deliverable Submission Date	30.12.2024
Task Leader/Main Author	INSAIT
Contributing Partners	ST, UEDIN, SINTEF, UNIMORE, THALES, SED, HES-SO, VICOM, CEA, CETIC, CSEM, DFKI, DLR, Fraunhofer, UCLM, BCA, INRIA
Reviewer(s)	Alain Pagani and Mohamed Selim

Keywords

Edge AI, Cooperative Inference Systems, Federated Learning, On-device Training, Spiking Neural Networks, Low-power AI, Neural Network Compression, Multi-task Learning, Event Cameras, Gesture Recognition, Model Compression, Continual Learning, DNN Acceleration, Unsupervised Learning, Adaptive Inference, Generative Models, Neuromorphic Computing

Revisions

Version	Submission date	Comments	Author
V0.1	21/12/2024	Algorithms and methods prototype – review version	INSAIT
V1.0	27/12/2024	Final version	DFKI

Disclaimer

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Acronyms and definitions

Acronym	Meaning
ADLIF	Adaptive Leaky Integrate-and-Fire
ARIMA	Autoregressive Integrated Moving Average
BCA	BONSEYES COMMUNITY ASSOCIATION
CEA	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES
CETIC	Centre d'Excellence en Technologies de l'Information et de la Communication
CIS	Cooperative Inference Systems
CL	Continual Learning
CNN	Convolutional Neural Network
COLAS	Continual Learning Across Scales
COTS	Commercial Off-The-Shelf
CSEM	CSEM CENTRE SUISSE D'ELECTRONIQUE ET DE MICROTECHNIQUE SA - RECHERCHE ET DEVELOPPEMENT
DAI	Distributed Artificial Intelligence
DFKI	DEUTSCHES FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH
DLR	DEUTSCHES ZENTRUM FUR LUFT - UND RAUMFAHRT EV
DNN	Deep Neural Network
DPU	Deep Learning Processing Unit
DSCT	Deadline Scheduling with Compressible Tasks
DSCT-EA	Deadline Scheduling with Compressible Tasks-Energy Aware
DSE	Design Space Exploration
DSP	Digital Signal Processor

FL	Federated Learning
FOMO	Faster Objects, More Objects
FPGA	Field-Programmable Gate Array
FRAUNHOFER	FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV
GAP	GreenWaves Technologies Architecture Processor
GMAC/s	Giga Multiply-Accumulate operations per second
GRU	Gated Recurrent Unit
HDR	High Dynamic Range
HES-SO	Haute École Spécialisée de Suisse Occidentale
HLS	High-Level Synthesis
HV	Horizontal-Vertical polarization
IoT	Internet of Things
LDA	Latent Dirichlet Allocation
LIF	Leaky Integrate-and-Fire
MLaaS	Machine Learning as a Service
MLOps	Machine Learning Operations
MSE	Mean Squared Error
MTL	Multi-Task Learning
NAS	Neural Architecture Search
ONNX	Open Neural Network Exchange
PEFT	Parameter-Efficient Fine-Tuning
PLIF	Parametric Leaky Integrate-and-Fire
PQ	Panoptic Quality
RISC-V	Reduced Instruction Set Computer - V

SAR	Synthetic Aperture Radar
SARIMA	Seasonal Autoregressive Integrated Moving Average
SED	SAFRAN ELECTRONICS & DEFENSE
SHD	Speech Commands High-Dimensional
SINTEF	SINTEF AS
SMPC	Secure Multi-Party Computation
SNN	Spiking Neural Network
SOD	Stage of ice Development
ST	STMicroelectronics
THALES	Thales Group
UCLM	UNIVERSIDAD DE CASTILLA - LA MANCHA
UEDIN	THE UNIVERSITY OF EDINBURGH
UMAP	Uniform Manifold Approximation and Projection
UNIMORE	Università degli Studi di Modena e Reggio Emilia
VICOM	FUNDACION CENTRO DE TECNOLOGIAS DE INTERACCION VISUAL Y COMUNICACIONES VICOMTECH
VRNN	Variational Recurrent Neural Network
YOLO	You Only Look Once

Executive Summary

This deliverable presents the first version of the edge AI algorithms, which serve to the objectives (O 5.1): Identify the key Edge AI hardware, middleware and software design and implementation challenges. Early versions of Edge AI algorithms and methods for resource-constrained devices have been developed to improve distributed learning strategies with continual, hierarchical and incremental learning. On the device level, neuromorphic softwares are developed for neuromorphic hardware. A major focus made is on deploying edge AI run-time inference with adaptation mechanisms in changing environments or conditions, considering multiple inputs (cameras, event cameras) and modalities (vision, audio, EEG) and tasks (2D segmentation, 3D reconstruction). A total of 18 patterns have directly participated in achieving objective 5.1, who have different expertise and collaborations targeted for specific or multiple usecases.

To achieve the objectives, the project has focused on integrating multidisciplinary approaches that bridge hardware limitations, algorithmic efficiency, and application-driven innovation. The development of edge AI algorithms has been organized into thematic subgroups to address the diverse challenges presented by resource-constrained environments. These efforts are made by dividing the sub-tasks of developing edge AI algorithms and methods into four sub-groups of: federated systems, continual learning, neuromorphic computing, and on-the-edge inference.

Federated systems: Federated systems within the dAIEdge framework enable collaborative inference and distributed learning across resource-constrained devices, optimizing performance while preserving data privacy and reducing latency. INRIA developed a novel Federated Learning (FL) framework for Cooperative Inference Systems (CISs), addressing client heterogeneity and surpassing state-of-the-art training methods with rigorous theoretical guarantees. CETIC achieved energy-efficient forecasting on Raspberry Pi devices by compressing models to a fifth of their size without accuracy loss. Both efforts emphasize Secure Multi-Party Computation (SMPC) and scalable FL architectures, advancing applications in multiple use cases.

Continual learning: Continual learning within dAIEdge focuses on developing adaptive edge AI systems capable of handling evolving data streams and dynamic environments. CEA has led collaborative efforts with partners like Ubotica, VERSES, and HIPERT to address edge learning challenges across domains, such as satellite vision, warehouse management, and smart cities. Achievements include adaptive triage systems for minimizing data drift, virtual lab platforms for benchmarking, and active inference models. Sparse label access methods and on-device retraining for personalized learning have been demonstrated using NVIDIA Jetson (by HSE-SO) and HAILO-8 (by UNIMORE) accelerators. Collaborative contributions include VICOMTECH's self-supervised learning for robotics, UEDIN's frameworks for accelerated online learning, and DLAS's across-stack co-design optimization for efficient deployment. Automated architecture search frameworks like *einspace* have also advanced resource-efficient neural network design for edge AI. These approaches enable scalable and robust edge learning solutions across use cases, including smart

city, space, and warehouse.

Neuromorphic systems: Neuromorphic systems are crucial for developing energy-efficient, real-time processing solutions in edge AI applications. CSEM has developed a Keras-based framework that enables the training of Spiking Neural Networks (SNNs) with learned connection delays, improving temporal precision for event-based tasks like audio classification. This framework supports various spiking neuron types and integrates with backpropagation. Additionally, CSEM is exploring how learned delays can optimize SNNs for image-based tasks. The goal is to create hybrid architectures combining spiking and non-spiking neurons, with the open-source library further promoting neuromorphic computing adoption. Meanwhile, Fraunhofer is focusing on event-camera-based 6D pose estimation, developing a synthetic training pipeline for efficient pose detection on edge devices, with collaborations in smart city and warehouse use cases. Note that the event cameras are neuromorphic by design.

On-the-edge inference: On-the-edge inference refers to running machine learning models directly on resource-constrained devices to minimize latency and bandwidth usage. Several advancements have been made in algorithms for this, such as Multi-task Learning (MTL) to share features across tasks, semantic segmentation with DINO-v2 for reduced memory footprint, and efficient sea-ice classification using unsupervised models with uncertainty quantification. Methods like saliency detection, gesture recognition with the FOMO algorithm, and lightweight audio signal processing models for edge devices have been developed to optimize performance with minimal resources. These innovations are being integrated in all three use cases of smart city, space, and warehouse. The developed method, e.g. edge AI for gesture and audio recognition, ensuring both efficiency and scalability in real-world scenarios.

Collaborative Frameworks and Use Case Integration

The success of these advancements is underpinned by collaborative efforts involving 18 partners, each contributing domain-specific expertise. Partnerships have facilitated the integration of developed algorithms into real-world use cases, such as:

- **Smart Cities:** Deployment of multi-task learning systems for efficient urban management, including parking slot monitoring and smart surveillance.
- **Space Applications:** Implementation of resource-efficient learning systems for sea-ice classification and satellite-based object detection.
- **Warehouse Applications:** Online self-supervised learning for robotic operations, minimizing in-hand errors and cycle times.

Throughout the project, the following involvements are prioritized, as per the original grant agreement. Due to the nature of the research, some level of flexibility is allowed at the same time. The leader of task 5.1 is INSAIT. This document also provides a detailed summary of the contribution of each partner in terms of research study as well as the usecase integrations.

Partner	Involvement
INSAIT (TL)	Study Multi-Task Learning (MTL) to lower memory footprint and computation time by feature sharing among tasks.
CEA	Develop active/self/semi/continuous learning methods for edge learning from a stream of unseen data with sparse access to new labels.
CETIC	Propose and demonstrate federated learning methods to enhance privacy and minimize latency of security services at the edge.
CSEM	Develop algorithms for high dynamic range ultra-low power image sensors and event-based imagers for low-power SNN and ONN architectures.
DFKI	Develop tools and services for deploying Neural Network models on low-energy devices including FPGA.
DLR	Identify and develop AI for classifications on satellites (space use case and T6.2).
FRAUNHOFER	Improve existing algorithms for 6D-Pose estimation using event cameras.
HES-SO	Contribute to the adaptation and integration of edge AI run-time inference and on-device training frameworks.
INRIA	Develop a network of devices that collaborate to provide inferences via personalized machine learning models.
SED	Set up distributed agreement, problem-solving, and optimization framework for edge AI using protocols and MiniZinc-based DCOP schema.
SINTEF	Investigate different edge AI platforms for implementing various AI algorithms on resource-constrained devices.
ST	Validate AI algorithms on resource-constrained general-purpose platforms, potentially with AI hardware accelerators.
THALES	Develop embedded ML algorithms for signal processing (audio, EEG, etc.) at the edge, focusing on power consumption optimization.
UCLM	Work on saliency detection for large data acquired in remote locations and efficient network updating with limited bandwidth.
UEDIN	Research Independent Efficient Edge-Learning Strategies beyond gradient-based methods, focusing on local online adaptation and communication.
UNIMORE	Study optimization techniques (e.g., mixed quantized training) for efficient continual learning on edge and resource-constrained devices.
VICOM	Develop Edge AI algorithms for inference and continual learning related to warehouse and space use cases.
BCA	Contribute to Edge AI run-time inference techniques.

CONTENTS

1. Algorithms and methods: Concept and Scope	11
1.1. Federated Systems.....	11
1.2. Continual Learning	12
1.3. Neuromorphic Systems.....	12
1.4. On-the-edge Inference.....	13
1.5. Contributions to use cases	13
2. Thematic Subgroups	14
2.1. Federated Systems.....	14
2.1.1. Cooperative Inference	14
2.1.2. Local Online Adaptation.....	15
2.2. Continual/Incremental Learning	16
2.2.1. Sparse Label Access.....	18
2.2.2. Personalized Edge Learning	18
2.2.3. Learning using Accelerators	19
2.2.4. Robotic Pick and Place	20
2.2.5. Human-Robot Continual Adaptation	20
2.2.6. Accelerated and Online Learning.....	20
2.3. Neuromorphic Systems.....	22
2.3.1. Algorithms for HDR Sensors.....	22
2.3.2. Event-Based SSNs.....	23
2.3.3. Event-Camera for Pose	24
2.4. On-the-edge Inference.....	25
2.4.1. Multi-task Learning	25
2.4.2. Satellite Image Understanding.....	26
2.4.2.1 Unsupervised models with uncertainty quantification	26
2.4.2.2 Supervised models	27
2.4.3. Saliency Detection.....	28

2.4.4.	Gesture Recognition	28
2.4.5.	Audio Signal Processing	30
2.4.6.	Bandwidth Efficient Update	31
2.4.7.	Efficient FPGAs Deployment	31
2.4.8.	Compressible Inference	33
2.4.9.	Inference on DPU-based SoC	34
3.	Conclusion and Future Work	35
3.1.	Problem Identification	35
3.2.	Future Works: Preparation for T5.2	35
4.	References	37

1. Algorithms and methods: Concept and Scope

This deliverable focuses on identifying the key design and implementation challenges in Edge AI hardware, middleware, and software. These challenges are explored and addressed through algorithms and methods developed within the project's context, as depicted in Figure 1. The project emphasizes four thematic sub-groups: Federated Systems, Continual Learning, Neuromorphic Systems, and On-the-Edge Inference, which collectively contribute to achieving the project's overarching goals.

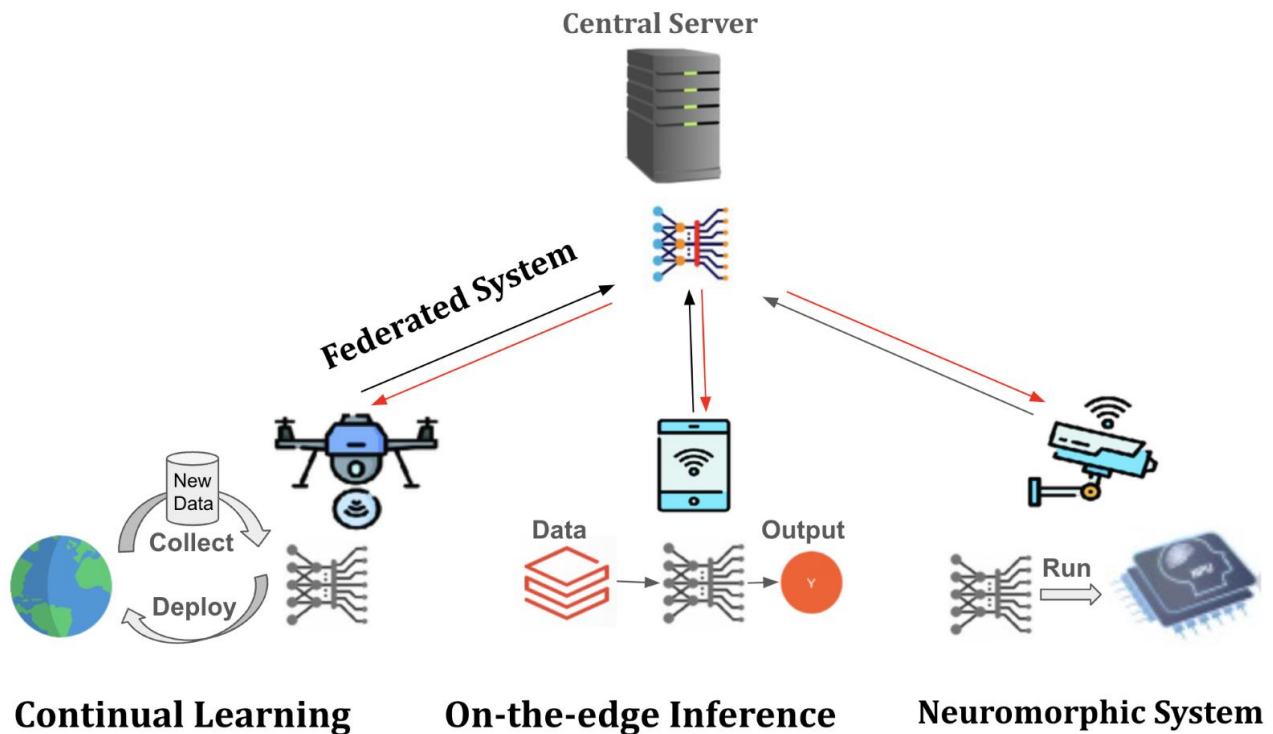


Figure 1: The system overview showcasing the relationships between different algorithms and methods studied within the scope of this project. This document is structured as per the sub-groups divided accordingly.

1.1. Federated Systems

Purpose and Contributions: Federated systems address the challenges of distributed learning by enabling collaborative inference and model training across resource-constrained devices while ensuring data privacy and reducing latency [1,2]. These systems leverage federated learning frameworks and hierarchical model designs to optimize edge AI performance. Key contributions include:

- **Collaborative Inference Systems (CIS):** A framework by INRIA that enhances model sharing across heterogeneous devices with minimal latency.
- **Efficient Forecasting:** CETIC developed lightweight models using compression techniques for deployment on Raspberry Pi devices, balancing the trade-offs between model accuracy, computational efficiency, and suitability for embedded systems,

ensuring that these models remain effective for real-world energy forecasting applications on small-scale devices.

- **Secure Multi-Party Computation (SMPC):** the system is secured using Secure Multi-Party Computation (SMPC), ensuring that data privacy and model integrity are maintained throughout the process.

Relevance to WP5: Federated systems contribute to WP5's objectives by addressing technical challenges (O5.1), fostering collaborations among Edge AI stakeholders (O5.2), and advancing privacy-focused, scalable architectures.

1.2. Continual Learning

Purpose and Contributions: Continual learning focuses on adaptive edge AI systems that handle evolving data streams and dynamic conditions [3,4]. It emphasizes personalized, semi-supervised, and sparse label learning, with examples including:

- **Identification of Needs and Challenges:** CEA has identified several needs and challenges of continual learning in all three different usecases.
- **Sparse Label Access:** Solutions address label scarcity in object detection pipelines, particularly in scenarios like satellite vision and smart city applications.
- **Personalized and Accelerated Learning:** Leveraging AI inference accelerators like HAILO-8, on-device training is optimized for resource-constrained hardware, enabling continual learning and real-time model adaptation. Collaborative efforts focus on integrating advanced learning methods, such as self-supervised robotic learning and efficient architecture search, to enhance edge AI applications across smart cities and dynamic industrial settings.

Relevance to WP5: Continual learning aligns with WP5's goal to map technological advances (O5.4) and ensures edge AI adaptability, contributing to robust and scalable solutions for dynamic environments.

1.3. Neuromorphic Systems

Purpose and Contributions: Neuromorphic systems play a pivotal role in developing energy-efficient, real-time processing solutions for resource-constrained devices [5]. Contributions include:

- **Spiking Neural Networks (SNNs):** CSEM developed frameworks to optimize spike transmission delays in SNNs, enhancing temporal precision in event-based tasks like audio classification.
- **Event Cameras for Pose Estimation:** Fraunhofer leveraged event-driven architectures for low-power 6D pose estimation, supporting smart city and warehouse use cases.
- **Hybrid Architectures:** Integration of spiking and non-spiking neurons for improved neuromorphic computing adoption.

Relevance to WP5: Neuromorphic systems address WP5's focus on hardware-software integration gaps (O5.3) and energy-efficient AI designs, aligning with the project's sustainability goals.

1.4. On-the-edge Inference

Purpose and Contributions: On-the-edge inference ensures real-time AI capabilities by running models directly on devices, reducing latency and bandwidth requirements [6]. Highlights include:

- **Multi-Task Learning (MTL):** INSAIT introduced a generalist framework using DINO-v2 for panoptic segmentation, depth prediction and object detection, lowering computation time and memory footprint, by performing multiple tasks at once.
- **Sea-Ice Classification:** DLR applied unsupervised learning with uncertainty quantification to classify sea ice in SAR satellite images, showcasing edge AI's versatility.
- **Gesture Recognition and Audio Signal Processing:** Systems for gesture recognition using the Edge Impulse platform's FOMO and efficient audio processing models were developed for micro-edge devices.

Relevance to WP5: On-the-edge inference supports WP5's goal to ensure secure AI deployment (O5.5) while demonstrating the potential of edge AI in diverse applications like smart cities and space exploration.

1.5. Contributions to use cases

The four sub-groups collectively address the core objectives of WP5 by providing holistic solutions that advance edge AI technologies across hardware, middleware, and software dimensions. These usecase relationships of algorithm and methods are shown in Figure 2. Some contributions that are integrated into key application domains:

- **Smart Cities:** Solutions like multi-task learning for urban management and event-camera-based systems for real-time surveillance highlight edge AI's potential in urban scenarios.
- **Space Applications:** Technologies such as sea-ice classification and efficient bandwidth updates demonstrate edge AI's role in satellite-based operations.

Warehouse Automation: Systems like self-supervised learning for robotics and gesture recognition for human-machine interaction exemplify edge AI's industrial relevance.

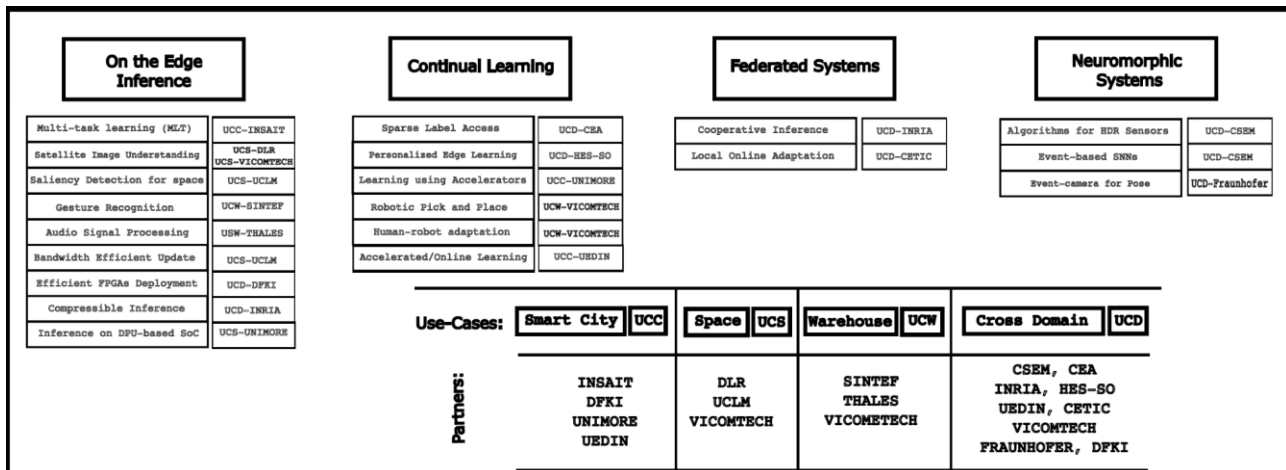


Figure 2: Four different sub-groups and sub-task per group. The correspondence to the use-cases and assigned partners are also given. Please refer to the bottom to establish the relationship between partners and the use cases.

In the following, more details of different sub-groups, individual partners' contributions, and their relationships to different usecases are provided. This document also provides the conclusion of the study performed to identify the key Edge AI design and implementation challenges, as a part of the objective O5.1.

2. Thematic Subgroups

2.1. Federated Systems

2.1.1. Cooperative Inference

Partner: INRIA

Task: Develop a network of devices that collaborate to provide inferences

Study: As the Internet of Things (IoT) technology advances, end devices like sensors and smartphones are progressively equipped with AI models tailored to their local memory and computational constraints. Local inference reduces communication costs and latency; however, these smaller models typically underperform compared to more sophisticated models deployed on edge servers or in the cloud. Cooperative Inference Systems (CISs) address this performance trade-off by enabling smaller devices to offload part of their inference tasks to more capable devices.

Developed Methods/Algorithms: CISs often deploy hierarchical models that share numerous parameters, exemplified by Deep Neural Networks (DNNs) that utilize strategies like early exits or ordered dropout. In such instances, Federated Learning (FL) may be employed to jointly train the models within a CIS. Yet, traditional training methods have overlooked the operational dynamics of CISs during inference, particularly the potential high heterogeneity in serving rates across clients. To address this gap, we propose a novel FL approach designed explicitly for use in CISs that accounts

for these variations in serving rates [10]. Our framework not only offers rigorous theoretical guarantees but also surpasses state-of-the-art (SOTA) training algorithms for CIs, especially in scenarios where inference request rates or data availability are uneven among clients. Moreover, our rigorous theoretical results are applicable to all approaches that jointly train models sharing a subset of parameters, including early exit networks, ordered dropout, pruning, and other nested training methodologies.

2.1.2. Local Online Adaptation

Partner: CETIC

Task: Compression techniques on energy forecasting models

Study: The main objective of this study is to assess the impact of model compression techniques on energy forecasting models, with the ultimate goal of making these models lightweight enough for deployment on resource-constrained devices like the Raspberry Pi. By reducing the model's complexity, we aim to maintain or even improve forecasting accuracy while ensuring that the model can run efficiently on smaller hardware. The second objective is to implement a federated learning solution across multiple Raspberry Pi devices using compressed models based on the Flower framework and use the best security option to secure the infrastructure.

To achieve this, we explored a variety of clustering techniques to preprocess the data before training the models. These techniques were tested to enhance the learning process and improve the overall model performance. Additionally, we conducted a comparative analysis between traditional time series forecasting models such as ARIMA and SARIMA, and more modern architectures, including N-Hist and N-Beast. Our evaluation focuses on balancing the trade-offs between model accuracy, computational efficiency, and suitability for embedded systems, ensuring that these models remain effective for real-world energy forecasting applications on small-scale devices.

Method: We utilized a public Irish dataset containing electricity consumption data. Preprocessing was essential, and we chose to apply spectral clustering as our method. Spectral clustering reduces the dimensionality of the data before performing clustering in fewer dimensions, allowing for more efficient grouping. The UMAP-based projection revealed three distinct clusters within the data. For model compression, we applied techniques such as pruning, compilation, and quantization to optimize performance. These compression methods were implemented using the ONNX framework to achieve an efficient, lightweight model suitable for deployment on constrained devices.

Result: The preliminary results obtained with the N-Beats model show a substantial reduction in model size, with compression techniques enabling us to shrink the model by a factor of five. Despite this compression, the model's predictive performance remained consistent, as evaluated using standard metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE). These metrics allowed us to closely monitor any potential trade-offs between model size and forecasting accuracy, ensuring that the compressed model maintained a high level of performance while significantly reducing its computational footprint.

Table 1: The B-Beats models compress by a factor of five without compromising the results.

Model	Model Size (MB)	Inference Speed (ms)	MSE	MAE
Original	90.4	51.8	1.06	0.601
Pruning	73.2	2.66	1.09	0.599
Compile	73.2	2.48	1.09	0.599
Quantization	18.7	0.74	1.10	0.600

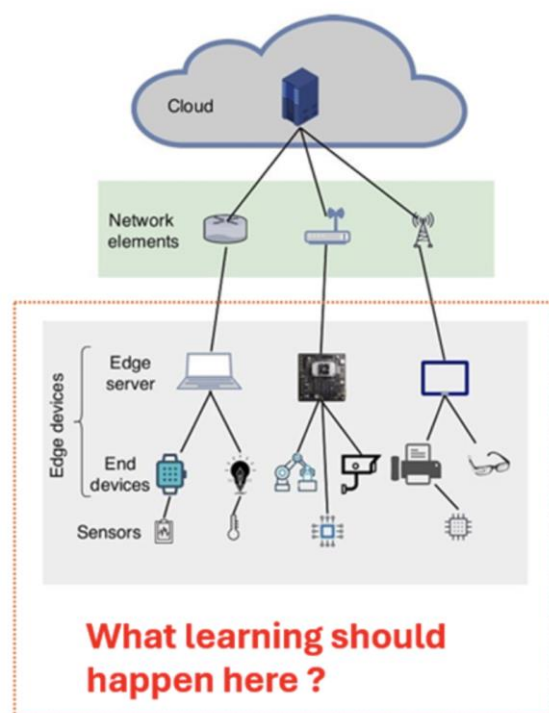
Conclusion: The results are very promising, with a model five times smaller while maintaining the same accuracy. The next step in this study will involve implementing a federated learning (FL) architecture across multiple Raspberry Pi devices using compressed models to optimize resources. Each Raspberry Pi will contribute by training the model locally, with results then aggregated centrally to update a global model. This approach will help preserve the privacy of local data, reduce bandwidth consumption, and tailor models to the limited capacities of the devices. Additionally, the infrastructure will be secured using Secure Multi-Party Computation (SMPC), ensuring that data privacy and model integrity are maintained throughout the process. In the context of CyberSecurity for AI on the edge we are working in collaboration with the dAIEDGE Federated Learning workgroup (consisting of CETIC, BTH, INRIA, KUL, USAL) in order to put together a secure federated learning architecture.

2.2. Continual/Incremental Learning

Partner: CEA

Task: Continuous edge learning (task-lead)

CEA has led this task (continuous edge learning) since October 2023. Regular meetings have been animated once per month with participation from a range of partners (CEA, FORTH, Ubotica, VERSES, CETIC, UoE, VicomTech, UoG, CSEM, HES-SO, INRIA, Hipert). The objective in the first phase of this task has been to align together on issues related to realistic learning systems at the edge (edge server, end devices, and near-sensor) and study the issues related to edge learning of the use case providing partners – Ubotica (on-satellite computer vision), VERSES (warehouse management) and Hipert (smart city).



As of October 2024, the first phase of the task is completed and will now focus on more targeted studies between partners to address issues identified in the first phase.

1. Improvement and in-depth study of Ubotica triage system for edge learning.

— It was found that there is a need for a solution where the model identifies what data it needs labeled and then uses this labeled data as effectively as possible to minimize drift and learn new classes. It has also been discussed that there is a need to identify a partner for Ubotica. It Perhaps is an interest of Hipert too. Additionally, it could be linked with ongoing work between VicomTech and Ubotica.

2. Use of a virtual lab platform to set up an edge learning benchmark node by HES-SO.

— Need to determine what such a benchmark should consist of.

3. Continue to follow ongoing work between VERSES and VictomTech on imitation learning.

— This is an interesting approach to learn with local observations from a human teacher and is being studied in the context of the smart warehouse use case of VERSES.

4. Development of an automated validation and anonymization tool to reduce the burden of model updating for Hipert.

5. CEA and VERSES have started exploring the development of adaptive active inference applications.

— CEA will also consider future efficient hardware implementations of active inference with respect to these use cases.

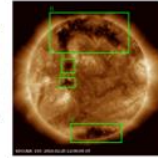
6. CEA is working with KU Leuven to understand the legal framework for edge learning.

— How can systems that change their operation autonomously in a distributed way at the edge be verified legally?

The second phase of the 5.1.2 task will also be led by CEA and will consist in encouraging progress in these five identified works that could be performed in order to advance continuous edge learning capabilities of the use case providers. The edge learning group meeting frequency will be diminished from monthly to perhaps quarterly and will serve principally to share progress in these tasks and no longer for open discussion as in the first phase.

Ubotica – satellite-based object detection

Need: data distribution shifts in time and new objects appear
Challenge: limited bandwidth to ground station (kB/s) to send data and receive model updates. How to learn locally ?

**VERSES****VERSES – adaptive warehouse management**

Need: warehouse environment is constantly changing
Challenge: adaptive intelligent agents (i.e., drones) path planning, object recognition and interaction with humans

Hipert – smart city real-time awareness

Need: privacy is a headache to update detection models
Challenge: to develop a more automated model life-cycle with model verification that is less human-intensive

**HIPERT**

Figure 3: Needs and challenges identified for the continual learning in different usecases.

2.2.1. Sparse Label Access

All three of these use cases converged to similar requirements and the same issues since they were largely based on object detection pipelines. The recurring theme was that, since the learning pipelines required some extent of supervision for fine-tuning and learning of new objects classes, the limiting factor was access to labels locally. In each of the three cases, it would be required to send data samples to a human labelling service and receive the label back at the edge in order to perform learning. We noted that some applications exist where labels might be obtained locally. For example, optical flow learning would require waiting some time to compute flow between successive frames. Otherwise, in car safety systems a “hands-on-wheel” detector could use rotation of the wheel as a labelling mechanism (although the data would be biased). In some applications, for example embedded health monitoring, labels may be obtained by posing questions to the user of the system. It was discussed at length whether, in the scenario where labels had to be obtained using a centralized human service, if edge learning made any sense at all. It was concluded that if a massively distributed learning scenario existed where the data distribution varies significantly for each deployed system it could be worthwhile (i.e., not needing to store many model weights centrally, being able to leverage unsupervised methods locally). In scenarios where bandwidth is severely limited (i.e., to send images and receive annotations), notably on a satellite, it may be more practical to send data instead of weights over the connections.

2.2.2. Personalized Edge Learning

Partner: HES-SO

Task: Edge learning

Study: This work explores the concept of retraining a pre-trained model on an NVIDIA device (Jetson Orin Nano), aiming to adapt this model to the specific needs of the user. The study focuses on the capabilities and limitations of on-device learning, a method that allows training to be performed directly on a device without relying on external resources [16,17,18].

The model used in this study was initially designed to detect the position of a driver's hands to determine if they are on the steering wheel. It is integrated into a solution developed by the company BEEemotion, which monitors the number of people in the car, their gender, emotions, and body positions, including hand positions. The goal was to adapt this model to be more personalized for each driver. To achieve this, a 'Retrain' button was proposed to be integrated into the solution, prompting the driver to perform specific actions, such as driving with only the left hand on the wheel, then the right hand, and so on, to build a new dataset. This dataset would then be used to retrain the model, better adapting it to the driver.

The main issue addressed in this work relates to the constraints that accompany on-device learning, including limited memory, restricted computational power, and resource management. These limitations present significant challenges for training complex models on the device. Several critical steps were taken to conduct this study. A review of existing literature and previous work in this field was conducted. Specific scripts were then developed for the various training phases. Multiple experiments were carried out to assess the feasibility of retraining multiple layers of the model and the impact of different hyperparameters. Multiple metrics, such as training time and memory allocation, were retrieved to assess the feasibility of the on-device training algorithm onto the Jetson Orin Nano by applying Onnx runtime training.

The main results obtained indicate that on-device retraining is indeed possible. Tests revealed a significant improvement in model accuracy after retraining, indicating successful adaptation to the driver.

2.2.3. Learning using Accelerators

Partner: UNIMORE

Task: Exploiting COTS AI inference accelerators for on-device training.

Study: On-device training is essential for enabling practical continual learning (CL) in real-world applications, facilitating life-long model adaptation to dynamic data streams by fine-tuning pre-trained models directly on resource-constrained hardware. Moreover, traditional AI co-processors are designed to accelerate deep learning inference. This research uses commercial off-the-shelf AI inference accelerators (e.g., HAILO-8) to empower on-device learning capabilities at the edge.

Developed Methods/Algorithms: To effectively deploy on-device training algorithms at the edge several works [38] [39] avoid performing complete updates and backpropagation. Still, they typically rely on a transfer-learning scheme or partial parameters update. In this context, our key idea is to rely on AI inference accelerators like HAILO-8 for accelerating the frozen graph of the model, keeping on the CPU the execution of only the part of the graph that performs the real backpropagation.

Use Case integrations and collaboration: UNIMORE, in collaboration with HIPERT, will study the usage of HAILO-8 accelerators and on-device training technology for the next generation of smart cameras for the smart city.

2.2.4. Robotic Pick and Place

Partner: VICOMTECH

Task: Study Online self-supervised learning for accurate robotic pick-place

Study: In modern industrial automation, a significant challenge is addressing in-hand errors, discrepancies between the expected and actual position of an object in a robotic gripper during vision-guided pick operations [43]. Flexible peg-in-hole strategies that can absorb residual in-hand errors have been extensively studied [44] to face this problem.

Our contribution proposes an online self-supervised learning method that aims to reduce cycle times by minimizing the need for corrective processes. We begin with the assumption that flexible peg-in-hole strategies can perform the assembly effectively but may increase cycle times. In an online manner, our system develops a regression model that learns to correct in-hand errors based on insights gained from corrective strategies. The system also evaluates the accuracy of the regression model in real-time. As new data becomes available, the system continuously retrains the model. The developed algorithm can decide when it can use the model that optimizes the process and when it should revert to traditional methods.

2.2.5. Human-Robot Continual Adaptation

Partner: VICOMTECH

Task: Study generative model in human robot cooperative task and its inference at the edge

Study: Recent advancements in generative models for human-robot cooperative tasks focus on task execution through predictive modelling and learning from human demonstrations. A notable approach [46] involves using a Variational Recurrent Neural Network (VRNN) to model the trajectory variations of human-robot teams over time. Moreover, Diffusion Policy [47] leverages the power of denoising diffusion probabilistic models (DDPMs) to train robot policies. This approach has shown its outstanding performance in various tasks where continuous adaptation and coordination are required.

We started with data capturing with a human-robot co-carry simulator provided in [46], where a human-human demonstration dataset is captured with joystick controllers, covering a range of initial conditions, goal conditions, and maps. An initial VRNN model is trained and validated with the recorded data.

2.2.6. Accelerated and Online Learning

Partner: UEDIN

Task: Accelerated and online learning

Study: The problem of being able to do learning in edge settings has a variety of particular specific concerns. These include (a) being able to learn from a data stream, (b) being able to cope with distribution shift (c) jointly taking account of the deployment device, architecture as well as machine learning model, and (d) choosing efficient architectures. This has been the subject of our

contribution to dAIEDGE in terms of research direction, research collaboration and network building. At this stage the primary collaboration has been with colleagues in UGLAS, but that is in the process of expanding.

Considering (a) the problem of learning from a data stream, we have shown [48] that the process of learning in an online fashion itself creates substantial model training failure, independent of any domain shift. We design the “chunking setting” as a benchmarkable setting to forward capability in this area and show that simple approaches can help improve learning in the chunking setting, demonstrating there is substantial headroom for others to develop capability that goes beyond this. We presented this work at COLAS 2024 in Pisa – the primary continual learning conference. Improvements in being able to reliably learn from streaming data are critical in the edge setting. This is being structured as a community benchmark.

We also consider the setting of distribution shift where we need to adapt a model to new previously unseen classes without forgetting old classes. We demonstrate a Bayesian learning approach [51] that can determine both a highly performant memory buffer and an effective use of that buffer for future stabilized adaptation. We demonstrate state of the art continual learning results in all the key settings.

Working with colleagues in UGLAS we have provided a conceptual framework for practitioners in AI at the edge. This conceptual model (DLAS [49]) for a cross-stack deep learning acceleration gives recipes for optimizing on-device methods for deployment. The methods discussed are suitable for both inference-time accelerations (with more matrix-vector computations) and learning-time accelerations (dominated by matrix-matrix operations). The DLAS model effectively balances simplicity and expressiveness, aiding experts from multiple fields in addressing co-design acceleration challenges. By conducting an across-stack perturbation study, we demonstrate the interdependence of DLAS layers, highlighting the importance of co-design. This study was performed using a specialized tensor compiler to conduct experiments on various combinations of parameters within the DLAS layers. We evaluated the effects on inference time and accuracy by altering DLAS parameters across two datasets, seven established DNN architectures, four compression methods, three types of algorithmic primitives (both sparse and dense), both untuned and auto-scheduled code generation, and four hardware platforms. Results showed significant variance in the performance outcomes based on the introduction of new DLAS parameters, such as changes in the performance ranking of algorithmic primitives dependent on the level of quantization. The work demonstrated the inherent value of co-design, but given this need coupled with the high costs associated with design space exploration (DSE), a conceptual framework such as DLAS is vital as a valuable conceptual framework for enhancing exploration of advanced accelerated deep learning solutions.

Finding excellent neural architectures for efficient machine learning is a challenge, and the value of architecture search in the edge domain makes automated neural architecture search an important tool. However, NAS search spaces are unhelpful and rigid, and are insufficiently expressive for either

novel architectures or highly efficient scenarios. Search spaces should be built upon more fundamental operations. To address this, we introduce einspace, a search space formulated around a parameterized probabilistic context-free grammar. This framework is adaptable, accommodating architectures of different sizes and complexities. It also supports a range of network operations, enabling it to model both convolutional and attention mechanisms, among others. It includes numerous pre-existing competitive architectures and provides the versatility needed to explore new possibilities. By using this search space, we conduct experiments to discover innovative architectures as well as enhance existing ones across the diverse Unseen NAS datasets. Our results demonstrate that we can achieve competitive architectures from scratch and consistently make significant improvements when we start the search with robust baselines. We contend that this approach marks a significant step forward in creating a transformative NAS paradigm, where the expressiveness of the search space and strategic search initiation are crucial.

Use Case integrations and collaboration: The developed method will be integrated in the smart city use case. INSAIT is directly collaborating with UNIMORE and HIPERT with experts on transformative approaches for deep learning in resource-constrained scenarios. Being able to perform on-the-edge-inference using the DINO-v2 backbone will enable many visual downstream tasks, beyond the context of this project.

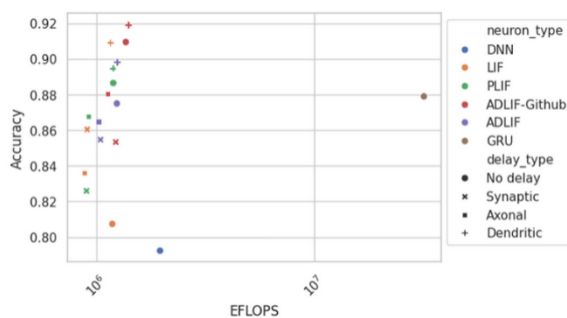
2.3. Neuromorphic Systems

2.3.1. Algorithms for HDR Sensors

Partner: CSEM

Task: Develop algorithms for high dynamic range ultra-low power image sensors and event-based imagers for low-power SNN and ONN architectures.

Study: In the field of edge AI, developing ultra-low power systems is critical for real-time processing in embedded devices. This effort involves both algorithmic development and sensor integration. High dynamic range (HDR) sensors and event-based imagers are increasingly used in applications that demand energy efficiency and rapid response times, such as robotics, wearable electronics, and medical imaging devices. Combining HDR technology with event-driven architectures can yield significant energy savings, further minimizing the power consumption of these systems.



Method	Neuron	Accuracy
Hammouamri et al.	(Conv) LIF + Synaptic delays	95.06%
Bittar et al.	AdLIF†	93.06%*
(Ours)	AdLIF† + Dend. Delay	92.13%
Sun et al.	SRM	92.24%
(Ours)	AdLIF†	92.00%
Deckers et al.	AdLIF†	91.67%
(Ours)	LIF + Dend. Delay	90.85%
(Ours)	AdLIF + Dend. Delay	90.19%
D'Agostino et al.	?	90.1%
(Ours)	PLIF + Dend. Delay	89.57%
Patino-Saucedo et al.	?	87%
Perez-Nieves	PLIF	82.7%

† use Github implementations * Maximum accuracy

Figure 4: 50x in EFLOPS compared to GRU models. Delays improve accuracy for equivalent power-budget

Table 2: Accuracy & computation for a network with 2x128 units, classifying audio digits of the SHD dataset

	Neuron	LIF	PLIF	ADLIF	DNN	GRU
	Delay					
Accuracy (%)	No delays	83.19	90.73	89.2	83.87	90.83
	Axonal	84.87	85.29	84.56	90.17	89.69
	Dendritic	90.31	91.3	89.31	91.17	-
	Synaptic	85.3	83.39	78.78	84.49	-
Effective FLOPs (K)	No delays	1194	1230	1545	2126	7435
	Axonal	920	940	1131	1732	7434
	Dendritic	1175	1194	1460	2021	-
	Synaptic	606	608	736	961	-

Developed Methods/Algorithms: To address these challenges, CSEM has developed a custom framework in Keras that facilitates the training of connection delays within Spiking Neural Networks (SNNs). This framework optimizes the timing of spike transmissions between neurons by allowing delays, which influence spike arrival times, to be learned alongside synaptic weights. The method is compatible with backpropagation, enabling the training of various types of spiking neurons. Our experiments show that learning delays significantly improve the temporal precision of SNNs in audio tasks, enhancing their capacity to handle event-based data streams efficiently while conserving energy. However, we have found that the specific type of delay plays a crucial role in determining the final accuracy. Our ongoing research is focused on identifying when different types of delays should be applied for optimal performance.

Use Case integrations and collaboration: To promote wider adoption, the developed library will be made open-source, aiming to democratize the use of SNNs within the Keras ecosystem. This initiative opens up the potential for hybrid architectures that integrate both spiking and non-spiking neurons, fostering new possibilities in neuromorphic computing and low-power AI applications.

2.3.2. Event-Based SSNs

A major challenge in designing edge AI low-power systems is optimizing energy usage without compromising performance. Recent advancements suggest that temporal data streams could benefit from learning delays within neural network architectures, which has already shown promise in event-based audio classification tasks [27]. However, these techniques have not yet been explored for temporal image data. In our study, we investigate how learned delays in spiking neural networks (SNNs) can enhance accuracy in image-based tasks, in comparison to traditional recurrent neural networks. Moreover, while most SNN libraries are designed to be compatible with PyTorch, there has been limited effort in developing robust implementations for Keras and TensorFlow.

2.3.3. Event-Camera for Pose

Partner: Fraunhofer IGD Darmstadt

Task: Improve existing algorithms for 6D-Pose estimation using event cameras.

Study: Processing high-framerate and high-resolution video streams at the edge demands significant power due to the large volume of data that must be handled, hindering the goal of low-power edge deployments. Event or neuromorphic cameras provide a low-power alternative to traditional frame-based vision [14], as they transmit and process only the changing areas of the frame. Considering pixels individually offers additional advantages, such as reduced latency and increased dynamic range, enabling deployment in more use cases. Surveillance is a particularly suitable application since the camera is typically stationary, only small parts of the image change, and outdoor scenarios often experience a high dynamic range. However, the sparse and asynchronous nature of event data poses a challenge for adapting classical vision algorithms and neural networks. While some progress has been made in 2D object detection, the task of 6D pose estimation remains largely unexplored.

Our goal is to develop an efficient neural network implementation for the edge that can perform 6D pose estimation on an embedded device. Additionally, the network should be trained using synthetic data, as generating 6D pose annotations for real data is impractical in most scenarios. By creating a fully synthetic training pipeline, we aim to serve the first and third use cases of WP6.

Developed Methods/Algorithms: We address the problem from two angles. First, we have created a labeled real dataset for 6D object pose estimation [11] that includes event data, which can be used to evaluate any proposed solutions. Second, we generate synthetic data [13] to automatically train and evaluate a 2D object detection network. This approach allows us to determine the optimal trade-off between precision and runtime on the target embedded platform. With a potentially usable network architecture in hand, we adapt it to actual event data by extending the synthetic data generation to simulate events [12] and assessing any performance implications of changing the data modality. In the final step, we will expand the 2D detection network to perform 6D pose estimation by elevating the 2D bounding boxes to 3D bounding boxes, enabling pose estimation [15].

Use Case integrations and collaboration: Fraunhofer is collaborating with HIPERT in the **Smart City** Use Case, where we are working on integrating our event-based pose estimation system with the embedded HIPERT hAURA platform. Furthermore, Fraunhofer is collaborating with VERSES on identifying and training a detection network in the **Warehouse Use Case**. The network is trained on fully synthetically generated data so it can be easily extended for new object categories.

2.4. On-the-edge Inference

2.4.1. Multi-task Learning

Partner: **INSAIT**

Task: Study Multi-Task Learning (MTL) to lower memory footprint and computation time by feature sharing among tasks.

Study: Across different use cases, the following six tasks of image classification, semantic segmentation, instance segmentation, depth prediction, object detection, object tracking were identified. The idea is to make a single forward pass to perform all tasks thus lowering the computation time. On the other hand, feature sharing among tasks lowers down the memory footprint which is essential for the edge devices. In the current stage, YOLO [31] based method is used to perform object detection, and the detected objects are tracked using the light-weight tracker that performs the object tracking by detection. The end-to-end object tracking requires the processing of the input videos, which is computationally expensive. Therefore, to meet the requirement of efficiency, we perform five different tasks by sharing the features, while the object tracking is performed using the existing light-weight tracker on the detected object obtained from the multi-task inference.

One key observation made is regarding the recent developments of vision foundation models, such as DINO-v2 [32], and CLIP [33]. These models can produce robust and generalizable visual features, by learning from a large collection of images, through self-supervised pre-training. These foundational models offer high performance for the many downstreamed visual tasks. However, harnessing the benefits of vision foundational models for the on-the-edge-inference MTL comes with two new challenges: (1) there exist a little to no study in the literature on how to leverage the foundational models for MTL; (2) the pre-trained vision foundational models are large in size, thus require high memory footprint. We address these two challenges for the on-the-edge MTL while enjoying the increased performance, robustness, and generalizability, thanks to the vision foundation models.

Developed Methods/Algorithms: We begin with two tasks: semantic segmentation, and panoptic segmentation which is also jointly known as panoptic segmentation. For this, we make use of the DINO-v2 foundational models as our vision encoder, as shown in the Figure below. Performance in Panoptic Quality (PQ) is shown in the table. Here, the proposed method achieves the best trade-off among performance, memory footprint, and computation time. More details of the developed method can be found here [34].

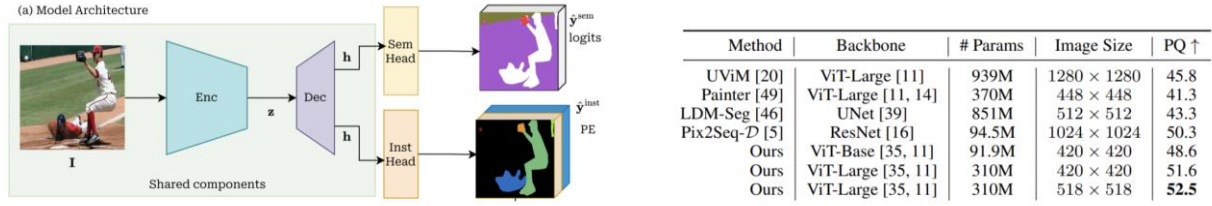


Figure 5: Multi-task generalist network evaluated for the task of panoptic segmentation.

We have extended the method in [4], for the task of object detection. At the same time, the ViT-based backbone is being further compressed and deployed in the NVIDIA Jetson Orin Nano edge device. The aim is to achieve objective detection performance similar to YOLO V11, while simultaneously performing four other tasks.

Use Case integrations and collaboration: The developed method will be integrated in the smart city use case. INSAIT is directly collaborating with UNIMORE and HIPERT with experts on transformative approaches for deep learning in resource-constrained scenarios. Being able to perform on-the-edge-inference using the DINO-v2 backbone will enable many visual downstream tasks, beyond the context of this project.

2.4.2. Satellite Image Understanding

2.4.2.1 Unsupervised models with uncertainty quantification

Partner: DLR

Task: Study Explainable Unsupervised Learning with uncertainty quantification for sea-ice classification task

Study: In this study, we focus on sea-ice classification with an unsupervised model with special importance on uncertainty quantification. The usefulness of unsupervised models comes with the following assumptions: 1) we have no or very little information (e.g. annotated data) available for training the model 2) Some domain expertise is available to understand the classified scenes obtained as output of the model

One critical problem in earth observation is the cost of labelling images. Although there are automatic labelling tools, human experts are needed in ascertaining the quality of the labelling. Unsupervised models with capabilities for uncertainty quantification can be used to reduce human effort in labelling yet maintain the quality of semantic segmentation by integrating domain knowledge in decision making. We have used a Bayesian hierarchical model called Latent Dirichlet Allocation (LDA) to discover latent data patterns in Synthetic Aperture Radar (SAR) Sea-ice images. Generative Probabilistic Model: LDA is an example of a Bayesian topic model, where, observations (e.g., words: in this case, clusters of SAR backscatter values) are collected into documents (in our case, SAR image patches) and each cluster's presence is attributable to one of the document's topics (in our case, sea-ice patterns). Each document will contain a small number of patterns.

Developed Methods/Algorithms: We address three issues of semantic segmentation here: 1) automatic discovery of data patterns without annotated information 2) interpretability 3)

Uncertainty quantization. LDA discovers patterns of information in the data, popularly known as ‘topics’ with regards to the original research [28]. Later, LDA was established as an explainable model in a sea-ice classification use-case [29][30]. Here we apply the workflow from [29] with an additional pipeline of uncertainty quantification.

Use Case integrations and collaboration: The developed method is integrated in the Space Use Case (sea-ice classification onboard a satellite from SAR data). Three sets of experiments are performed based on the combinations of polarization, i.e., HH, HH-HV, combination of HH, HV, and avg (HH+HV). The obtained results show a consistent structure and serve as a visual input for domain knowledge. In other words, the maps generated by LDA can be compared with some available domain knowledge and/or checked by a domain expert for assignment of semantic class names. LDA considers contextual information in the image scenes, while working on top of pixel-based clustering methods and hence, provide much more meaningful (in the context of human understanding of classes).

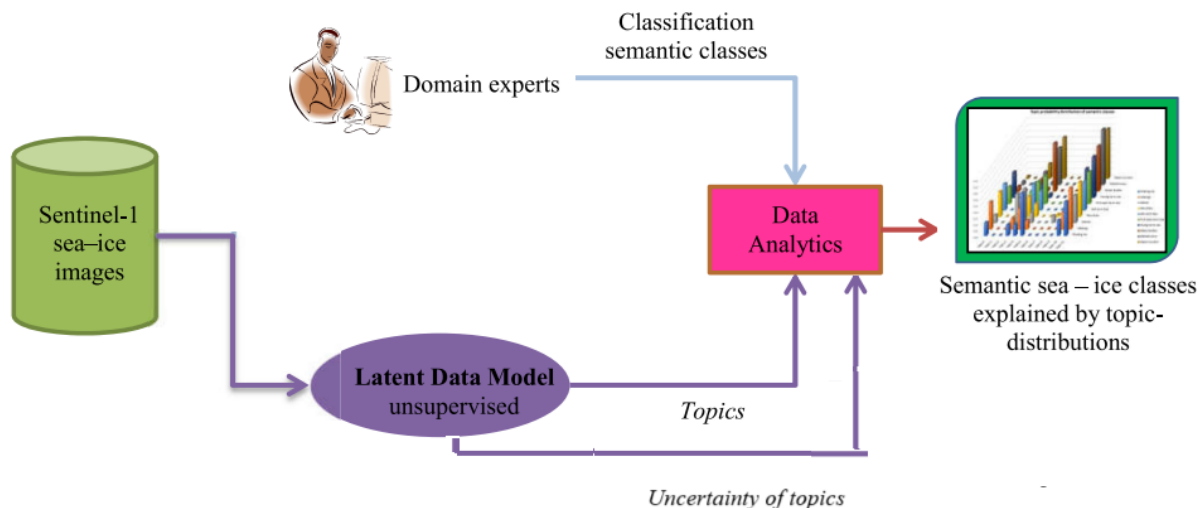


Figure 6: Workflow of LDA in sea-ice classification.

2.4.2.2 Supervised models

Partner: VICOMTECH

Task: Classify satellite images at the edge

Study: This study has focused first on applying and evaluating different exploratory, preprocessing, data augmentation and training strategies targeting sea-ice image classification. The objective is to train a model able to classify sea-ice satellite images on the edge with the requirements specified at the use case, including ones imposed by hardware restrictions. Secondly, the study has also targeted the analysis of MLOps architectures measuring data drifts to launch retraining processes to avoid model accuracy losses.

Use Case integrations and collaboration: The developed method will be integrated in the smart warehouse and the space use case. VICOM is directly collaborating with 1) VARJO with experts on

hand tracking algorithms to integrate richer observation in the model, and 2) DLR and partners from the space use case to further improve classification algorithms and analyze the suitability of the drift detection and retraining functionalities within the use case.

2.4.3. Saliency Detection

Partner: UCLM

Task: saliency detection, which is essential when vast amounts of data are acquired in remote locations

Study: We have decided to explore the ‘saliency detection’ (also called ‘anomaly detection’ in some contexts) approach in the problem of weapon detection from surveillance images. This application aligns with our work in work package 6, where our objective is to develop a weapon detector within the context of smart cities. For this, the approach is to consider that the normal situation is where people enter the building with either empty hands or common items in their hands (smartphones, keys, etc.). Objects such as handguns will be considered the ‘anomaly’ to be detected.

State-of-the-art weapon detectors leverage both visual information (like in standard object detectors) and body pose information (as the latter can be very informative for cases in which the handgun appears too small or against a dark background, this was first shown in [52]). Body pose information is generally obtained as a set of keypoint coordinates, corresponding to knees, elbows, neck, etc. We have started exploring one-class classifiers (also called anomaly detectors) on the keypoints, with the aim of detecting anomalies in the poses. When that is finished, we will proceed to apply the same approach with the visual information, and then combine both so as to compare with the state of the art.

2.4.4. Gesture Recognition

Partner: SINTEF

Task: Gesture recognition edge AI system based on RGB-camera input running on micro- and deep-edge devices.

Study: This study focuses on developing an edge AI algorithm for gesture recognition that leverages the Edge Impulse platform's FOMO (Faster Objects, More Objects) [19] machine learning algorithm. FOMO is designed for real-time multi-object detection and is suited for resource-constrained micro- and deep-edge devices. The goal is implementing a system that can recognize hand gestures in real-time using RGB camera input, enabling intuitive user interactions in various applications.

The study aims to demonstrate the potential of combining efficient multi-object detection with real-time performance on edge devices such as STMicroelectronics STM32MP157F-DK2 development kit. By leveraging the capabilities of FOMO, this system can accurately recognize and interpret hand gestures, paving the way for innovative applications.

The aim of the study is multi-folded, as described below:

- Utilize the FOMO algorithm for efficient multi-object detection for gesture recognition.
- Ensure the algorithm operates with minimal latency on edge devices with limited computational resources.
- To create a robust gesture recognition system that considers different lighting conditions and user scenarios.
- To design a lightweight gesture recognition model that can run efficiently on micro- and deep-edge devices.
- To leverage RGB camera input for capturing gesture data while minimizing energy and resource utilization and maintaining acceptable accuracy.
- To evaluate the performance of the minimizing diverse environmental conditions and user scenarios

Developed Methods/Algorithms: The workflow includes data collection using a diverse dataset of hand gestures captured using RGB cameras. The dataset consists of various gestures, angles, distances, and lighting conditions. The model used is the FOMO algorithm provided, which is optimized for real-time multi-object detection and is based on a lightweight convolutional neural network (CNN) architecture that balances performance and resource efficiency. The data preprocessing consists of preparing the RGB images to standardize input sizes and enhance the dataset through augmentation techniques. The input for the current architecture is an image sequence of 128x128 RGB images captured at 15 fps for a duration of 2s, and the net currently distinguishes between a motion in which the subject either moves their hand in an upwards or downwards motion. The camera used is a web camera with a resolution of 1080p.

The gestures selected are based on the standard gestures used in different domains (e.g., military, police, construction, etc.). For this study, the gestures were selected from the military field manual for gesture control FM 21-60 [20]. The gestures need to be recognized relatively easily both for the test set and during live inference, although further training would be required for it to be stable, which is the case for any set of gestures tested.

The workflow includes training the FOMO model using different datasets, configuring the model parameters, monitoring model performance during training, and optimizing the deployment on the STM32MP157F-DK2 device.

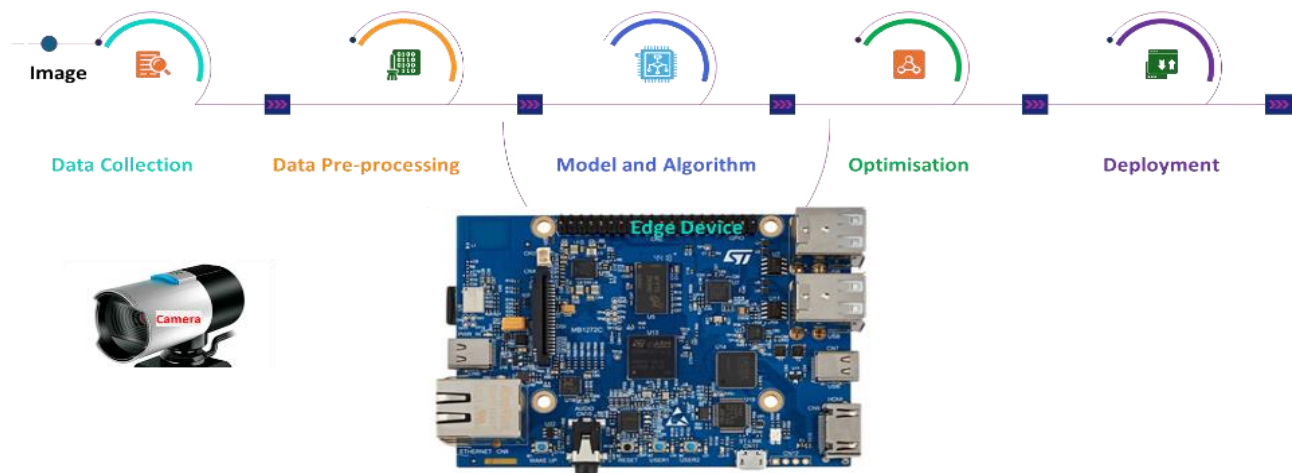


Figure 7: Gesture recognition edge AI system workflow.

Use Case integrations and collaboration: SINTEF is collaborating with ST to evaluate the different edge AI platforms for implementing various AI algorithms on edge resource-constrained devices. The study uses ST edge HW platforms and platforms enhanced with AI HW accelerators.

2.4.5. Audio Signal Processing

Partner: **THALES**

Task: Develop embeddable by design ML algorithms for signal processing (audio data)

Study: To promote multimodal data exploitation in dAIEDGE, we propose to enrich the algorithm library developed in the project (focused on image data) with other modalities, in particular, for Thales, with audio data.

Regarding hardware acceleration, many works have been done on image, much less on signal processing, especially on audio data. This type of data requires specific algorithms and architecture: 1D convolution and recurrent architecture..., transformers on audio data also emerge in the research community.

Another challenge concerns the deployment of these algorithms at the edge (in drones, for example, in the context of the smart warehouse use case), needing a hardware platform on top of which could be embedded and also run at very low consumption (< 50 mW) speech enhanced algorithms.

Developed Methods/Algorithms:

Current development focuses on two algorithms and their deployment on GAP9 devices.

GAP9 device presents very interesting characteristics for the development of audio processing:

- GAP9 processor | GreenWaves Technologies ([greenwaves-technologies.com](https://www.greenwaves-technologies.com))
- 8 cores,
- DSP,
- NN + signal processing accelerator, based on RISC-V architecture,
- <50 mW,
- 3.7mm x 3.7mm,
- 370 MHz internal clock.

Using this device, Thales currently works on two algorithms:

- Tiny Denoiser (1M parameters):

- Contains 3 linear layers and 2 GRUs
- 1M parameters and 0.20 GMAC/s: not yet reached the maximum capacity of GAP9: 2M and 0.6 GMAC/s, this opens a future work to generalize this algorithm
- Implemented and ran in real time on GAP9 with 25ms latency
- Public Models for audio data: MP-SENet (denoising) (much more parameters): [55]
 - Not yet implemented on GAP9: work in progress
 - Ran in real time but with a high latency (1s): work in progress

To develop and test the models, Thales uses public data: VoxCeleb and LibriSpeech.

Use Case integrations and collaboration: An application using multimodal data could be envisioned in the context of Smart Warehouse use case, enriching it with intrusion detection capabilities based on multimodal data including audio and image data. Thales proposed an open call highlighting this kind of work. Open call will be published in January 2025.

2.4.6. Bandwidth Efficient Update

Partner: UCLM

Task: Efficient updating of networks with limited bandwidth.

Study: As for the topic ‘efficient updating of networks with limited bandwidth’, this refers to the capacity of updating an AI model onboard a satellite with the minimum changes in the model weights. The model onboard has been previously trained on synthetic data, or with data captured from other sensors. Once the satellite is working, captures with the sensor onboard can be downlinked and used for retraining the original model. Then the updated model can be uplinked to the satellite, thus improving its performance. However, the changes with respect to the original model have to be minimal since the uplink bandwidth is generally extremely limited. Obviously, there is a trade-off between the number or magnitude of changes and performance gain. This is precisely what we should explore. We have selected reference work [53], that introduced a method based on selecting a set of weights to freeze, based on the magnitude of the numerical changes during training.

For this problem we intend to explore, among others, techniques from the so-called PEFT (Parameter-Efficient Fine-Tuning) approach [54], which has been studied in the context of efficient retraining and finetuning of LLMs. As opposed to [53], in this case we will use a dataset representative of the space scenario, like for example for ship detection or segmentation in satellite images.

2.4.7. Efficient FPGAs Deployment

Partner: DFKI

Task: Develop tools and services for deploying Neural Network models on low-energy devices including FPGA.

Study: Using GPUs at the edge is limited by their high-power requirements and large physical size, which makes them unsuitable for small, energy-constrained edge devices like sensors and cameras. GPUs, typically housed in data centers, are not practical for on-site processing where low latency and efficiency are critical. This is where FPGAs become essential, as they offer a more energy-efficient alternative that meets edge computing needs without sacrificing performance. FPGAs can be tailored for specific DL tasks, ensuring faster, more efficient processing closer to users while meeting the constraints of edge environments. However, a major challenge in using FPGAs for edge computing is fitting large neural networks onto the limited hardware resources of FPGAs while maintaining the model's accuracy and performance [21][22]. Our goal is to provide efficient neural network implementations for on-the-edge inference which serve the 1st and 2nd use cases of WP6.

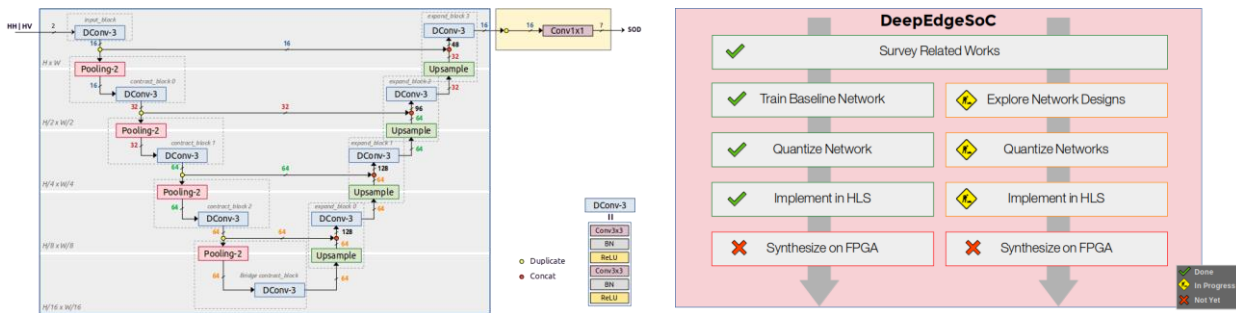


Figure 8: Sea Ice Classification on FPGA for Satellite On-Board Processing.

Developed Methods/Algorithms: As shown in the figure below, we begin with preparing the dataset related to the task or the use case. Afterwards, the network design space is explored by training different network topologies and choosing the most suitable one for the task in terms of inference speed, accuracy, and memory requirements. Once a network is adopted, it is quantized by using fixed point format instead of the conventional floating-point format. This helps reduce the memory demand as well as the computation and transfer time on the hardware. The network is then implemented in C++ for High-Level Synthesis (HLS) and eventually deployed on the FPGA System-on-Chip (SoC). The development steps are performed within the DeepEdgeSoC framework [21].

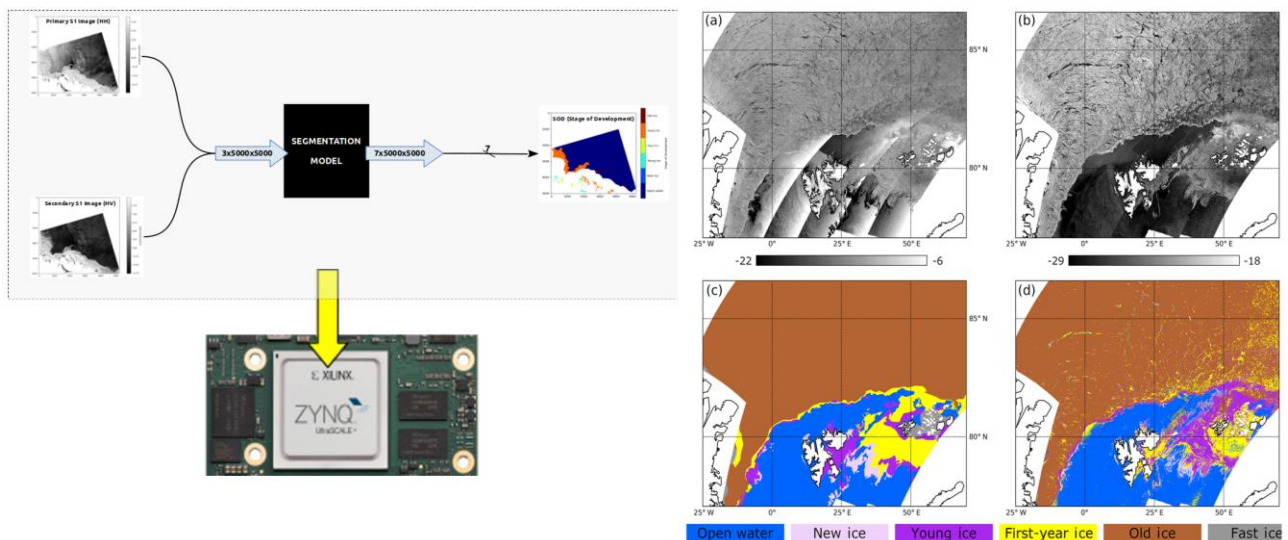


Figure 9: The sea-ice semantic segmentation for SAR S1 images quantized and deployed on the Xilinx XCZU9EG SoC

Use Case integrations and collaboration: DFKI is collaborating with UNIMORE in the **Smart City Use Case**, where we are working on integrating our generated IP Core with UNIMORE overlay on the FPGA. Furthermore, DFKI is collaborating with DLR on designing a sea ice classification network in the **Space Use Case**. The network outputs a Stage of ice Development (SOD) map based on Sentinel S1 SAR images (HH + HV). The AI4Arctic dataset [23] is used for this purpose.

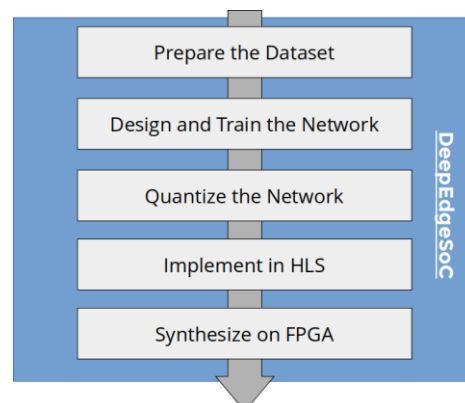


Figure 10: The Deep Edge Soc Pipeline.

2.4.8. Compressible Inference

Partner: INRIA

Task: Scheduling Machine Learning Compressible Inference Tasks

Study: With the advent and the growing usage of Machine Learning as a Service (MLaaS), cloud and network systems are now offering the possibility to deploy ML tasks on heterogeneous clusters. Then, network and cloud operators have to schedule these tasks, determining both when and on which devices to execute them. In parallel, several solutions, such as neural network compression, were proposed to build small models which can run on limited hardware. These solutions allow choosing the model size at inference time for any targeted processing time without having to re-train the network.

Developed Methods/Algorithms: To implement such solutions, the first task is to study how much ML models can be sparsified. We made two contributions on this matter. We first carried out experiments to extend the Once-For-All [35] solution for a full range of processing times and derived the full tradeoff between accuracy and processing time in [7]. Second, in [8] we provide new results on the Strong Lottery Ticket Hypothesis (SLTH) [36,37], stating that a random neural network N contains subnetworks capable of accurately approximating any given neural network that is sufficiently smaller than N , without any training. We provide the first proof of the SLTH in classical settings, such as dense and equivariant networks, with guarantees on the sparsity of the subnetworks.

The second task is to see how to schedule ML models. We considered the Deadline Scheduling with Compressible Tasks (DSCT) problem [7]: a novel scheduling problem with task deadlines where the tasks can be compressed. We also considered a variant, the Deadline Scheduling with Compressible Tasks-Energy Aware (DSCT-EA) problem [9], which addresses the scheduling of compressible machine learning tasks on several machines, with different speeds and energy efficiencies, under an energy budget constraint. For both problems, we propose approximation algorithms with proven guarantees to solve them and validate their efficiency with extensive simulation on deep learning classification jobs, achieving near-optimal results. Experimental results show that our approach allows us to save up to 70% of the energy budget of image classification tasks, while only losing 2% of accuracy compared to when not using compression.

2.4.9. Inference on DPU-based SoC

Partner: UNIMORE

Task: Efficient Edge AI inference on DPU-based SoC through Proxy Computing Paradigm

Study: FPGA-based heterogeneous systems are increasingly becoming a popular solution for the acceleration of Deep Neural Networks (DNNs). While mature design solutions simplify the deployment of the acceleration logic, the efficient integration and orchestration of several HW and SW tasks still remain a challenging and unsolved problem.

Developed Methods/Algorithms: Previous work explored the possibility of deploying part of the computation to the soft-core [40], [41]. Our methodology facilitates the definition and deployment of a cluster of accelerators that share local memory and a programmable core (the Proxy-Controller), both featuring tight coupling to the accelerators, DMAs, and other IPs in the cluster. The Proxy-Controller acts as an orchestrator for the accelerators inside the cluster, avoiding costly interactions between the host CPU and accelerators. As a representative embodiment of the paradigm, we are studying a system that leverages the commercial AMD/Xilinx Deep Learning Processing Unit (DPU) engine and a number of other accelerators executing layers that the DPU does not support.

Use Case integrations and collaboration: UNIMORE is collaborating with DFKI in this activity to integrate automated generation and accelerators for unsupported NN layers. The work will be evaluated by realistic applications and NN models in the smart-city domain.

UNIMORE is also developing and optimizing a novel end-to-end system architecture for a real-time video-based parking management system optimized for edge devices (e.g., Nvidia Orin). The proposed system integrates state-of-the-art object detection and tracking methods with novel Static Object Detection techniques to identify potential parking spaces. It also includes a separate parking occupancy classification pipeline that provides fast and accurate availability status of these parking spaces. The system is designed to minimize manual efforts and lower implementation costs, making it a practical solution for urban environments. UNIMORE, in collaboration with HIPERT, is working on the deployment of a Parking Slot Detector on the Haura devices to provide in real-time a suitable parking area to the biker of the smart city use case scenario.

3. Conclusion and Future Work

3.1. Problem Identification

This deliverable addresses the challenge of developing advanced edge AI technologies for resource-constrained environments. Key issues include ensuring efficient hardware-software integration, enabling real-time and adaptive inference, and enhancing distributed learning strategies across diverse application areas such as smart cities, space and warehouse applications. Additionally, the project seeks to promote energy efficiency, data privacy, and scalable solutions to foster Europe's competitive edge in edge AI innovation.

This project has so far identified several challenges within the scope of the usecases. In the space application, data distribution shifts over time, and new objects continuously emerge. The limited bandwidth to the ground station makes updating models on the ground impractical, necessitating efficient and continual on-device inference. In the smart city application, privacy concerns demand the implementation of federated learning, which not only facilitates inference on the edge but also supports continual learning. In the warehouse management setting, diverse edge devices with varying computational hardware are simultaneously utilized, requiring both learning and inference to be optimized to align with the architecture of the available hardware.

3.2. Future Works: Preparation for D5.2

Looking forward to the deliverable **D5.2 (Algorithms and methods final implementation)**, the project will continue focusing on the Objective O5.2: Develop solutions by leveraging synergies among the Edge AI stakeholders. This will be carried out in accordance with the T 5.2: Middleware and networks for edge AI [M04-M30] Map communication requirements with network performance criteria (reliability, delay, delay variation, throughput) and develop novel mapping of edge AI performance requirements and communication layer requirements. This will be aimed to serve Milestone 5: Final version of scientific and technical advances, of the work package 5: Edge AI Technological Advances for Cross-fertilization. In the current stage, below are a few examples of the identified future works:

1. **Optimizing Deployment:** Enhancing algorithms for deployment on diverse hardware platforms, including GPUs, FPGAs, and neuromorphic chips, to achieve maximum efficiency.
2. **Scalability Testing:** Expanding integration efforts to test scalability across larger networks of interconnected edge devices.
3. **Refinement and Validation:** Refining the developed solutions using real-world datasets and conducting extensive performance validation in dynamic and heterogeneous environments.
4. **Cross-Sector Innovation:** Strengthening synergies across thematic sub-groups to create hybrid systems that combine federated learning, continual adaptation, and neuromorphic computation.
5. **Security and Privacy Mechanisms:** Embedding robust data privacy and secure computation mechanisms tailored for distributed edge AI applications.

4. References

- [1] Zhang, Chen, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. "A survey on federated learning." *Knowledge-Based Systems* 216 (2021): 106775.
- [2] Pandya, Sharnil, Gautam Srivastava, Rutvij Jhaveri, M. Rajasekhara Babu, Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, Spyridon Mastorakis, Md Jalil Piran, and Thippa Reddy Gadekallu. "Federated learning for smart cities: A comprehensive survey." *Sustainable Energy Technologies and Assessments* 55 (2023): 102987.
- [3] Wang, Liyuan, Xingxing Zhang, Hang Su, and Jun Zhu. "A comprehensive survey of continual learning: theory, method and application." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [4] Van de Ven, Gido M., Tinne Tuytelaars, and Andreas S. Tolias. "Three types of incremental learning." *Nature Machine Intelligence* 4, no. 12 (2022): 1185-1197.
- [5] Roy, Kaushik, Akhilesh Jaiswal, and Priyadarshini Panda. "Towards spike-based machine intelligence with neuromorphic computing." *Nature* 575, no. 7784 (2019): 607-617.
- [6] Wang, X., Han, Y., Leung, V. C., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(2), 869-904.
- [7] Tiago da Silva Barros, Frédéric Giroire, Ramon Aparicio-Pardo, Stéphane Pérennes, Emanuele Natale. Scheduling with Fully Compressible Tasks: Application to Deep Learning Inference with Neural Network Compression. CCGRID 2024 - 24th IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing, IEEE/ACM, May 2024, Philadelphia, United States, 2024. [⟨10.1109/CCGrid59990.2024.00045⟩](#). [⟨hal-04497548⟩](#)
- [7] Tiago da Silva Barros, Frédéric Giroire, Ramon Aparicio-Pardo, Stéphane Pérennes, Emanuele Natale. Scheduling with Fully Compressible Tasks: Application to Deep Learning Inference with Neural Network Compression. CCGRID 2024 - 24th IEEE/ACM international Symposium on Cluster, Cloud and Internet Computing, IEEE/ACM, May 2024, Philadelphia, United States, 2024. [⟨10.1109/CCGrid59990.2024.00045⟩](#). [⟨hal-04497548⟩](#)
- [8] Emanuele Natale, Davide Ferré, Giordano Giambartolomei, Frédéric Giroire, Frederik Mallmann-Trenn. On the Sparsity of the Strong Lottery Ticket Hypothesis. In Proceedings of NeurIPS 2024, the Thirty-Eighth Annual Conference on Neural Information Processing Systems, Vancouver, Canada, Dec. 2024.
- [9] Tiago da Silva Barros, Frédéric Giroire, Ramon Aparicio-Pardo, Stéphane Pérennes. Scheduling Machine Learning Compressible Inference Tasks with Limited Energy Budget. In Proceedings of ICPP 2024, the 53rd ACM International Conference on Parallel Processing, ACM, Gotland, Sweden, Aug. 2024. [⟨10.1145/3673038.3673106⟩](#)
- [10] C. Kaplan, A. Rodio, Tareq Si Salem, Chuan Xu, Giovanni Neglia. Federated learning for cooperative inference systems: the case of early exit networks, [ArXiv:2405.04249](#)
- [11] Rojtberg, Pavel, and Thomas Pöllabauer. "YCB-Ev 1.1: Event-vision dataset for 6DoF object pose estimation." ECCV 2024 Workshops (2024).

- [12] Joubert, Damien, et al. "Event camera simulator improvements via characterized parameters." *Frontiers in Neuroscience* 15 (2021).
- [13] Hodaň, Tomáš, et al. "Photorealistic image synthesis for object instance detection." 2019 IEEE international conference on image processing (ICIP).
- [14] Hamann, Friedhelm, et al. "Low-power Continuous Remote Behavioral Localization with Event Cameras." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
- [15] Maji, Debapriya, et al. "YOLO-6D-Pose: Enhancing YOLO for Single-Stage Monocular Multi-Object 6D Pose Estimation." 2024 International Conference on 3D Vision (3DV). IEEE, 2024.
- [16] Microsoft, 2023. On-device training with ONNX runtime: a deep dive [online]. Microsoft Open Source Blog. URL: <https://opensource.microsoft.com/blog/2023/07/05/on-device-training-withonnx-runtime-a-deep-dive/>
- [17] TensorFlow Team, 2021. On-device training with TensorFlow Lite [online]. TensorFlow Blog. URL: <https://blog.tensorflow.org/2021/11/on-device-training-in-tensorflow-lite.html>
- [18] Kostadinov, Simeon, 2020. Understanding Backpropagation Algorithm [online]. Towards Data Science. URL: <https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd>
- [19] FOMO: Object detection for constrained devices. <https://docs.edgeimpulse.com/docs/edge-impulse-studio/learning-blocks/object-detection/fomo-object-detection-for-constrained-devices>.
- [20] Visual Signals Military Manual, 1987, FM 21-60.
[https://www1.radford.edu/content/dam/colleges/chbs/rotc/Forms/fm/Visual Signals FM 21-60.pdf](https://www1.radford.edu/content/dam/colleges/chbs/rotc/Forms/fm/Visual%20Signals%20FM%2021-60.pdf)
- [21] Al Koutayni, Mhd Rashed, Gerd Reis, and Didier Stricker. "Deepedgesoc: End-to-end deep learning framework for edge iot devices." *Internet of Things* 21 (2023): 100665.
- [22] Al Koutayni, Mhd Rashed, et al. "Real-time energy efficient hand pose estimation: A case study." *Sensors* 20.10 (2020): 2828.
- [23] J. Buus-Hinkler, "AI4Arctic Sea Ice Challenge Dataset". Technical University of Denmark, 21-Nov-2022, doi: 10.11583/DTU.c.6244065.v2.
- [24] Model Compression for Deep Neural Networks: A Survey, Zhuo Li et al., 2023, <https://www.mdpi.com/2073-431X/12/3/60>
- [25] N-BEATS: Neural basis expansion analysis for interpretable time series forecasting, Boris N. Ores et al., 2024, <https://arxiv.org/abs/1905.10437>
- [26] Flower: A Friendly Federated Learning Research Framework, Daniel J. Beutel et al, 2020, <https://arxiv.org/abs/2007.14390>
- [27] I. Hammouamri, I. Khalfaoui-Hassani and T. Masquelier, "Learning Delays in Spiking Neural Networks using Dilated Convolutions with Learnable Spacings," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [28] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
- [29] C. Karmakar, C. O. Dumitru, G. Schwarz and M. Datcu, "Feature-Free Explainable Data Mining in SAR Images Using Latent Dirichlet Allocation," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 676-689, 2021, doi: 10.1109/JSTARS.2020.3039012.

- [30] C. Karmakar, C. O. Dumitru, N. Hughes and M. Datcu, "A Visualization Framework for Unsupervised Analysis of Latent Structures in SAR Image Time Series," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5355-5373, 2023, doi: 10.1109/JSTARS.2023.3273122.
- [31] Redmon, J. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [32] Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez et al. "DINOv2: Learning Robust Visual Features without Supervision." *Transactions on Machine Learning Research*.
- [33] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PMLR, 2021.
- [34] Prasadnikov, Nedyalko, Wouter Van Gansbeke, Danda Pani Paudel, and Luc Van Gool. "A Simple and Generalist Approach for Panoptic Segmentation." *arXiv preprint arXiv:2408.16504* (2024).
- [35] H. Cai, C. Gan et al., "Once-for-all: Train one network and specialize it for efficient deployment," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=HylxE1HKwS>
- [36] Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *International Conference on Learning Representations*. Sept. 2018. (Visited on 10/20/2023).a
- [37] Vivek Ramanujan et al. "What's Hidden in a Randomly Weighted Neural Network?" In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020, pp. 11890–11899. DOI: 10.1109/CVPR42600.2020.01191.
- [38] Ravaglia, L., Rusci, M., Nadalini, D., Capotondi, A., Conti, F., & Benini, L. (2021). A tinyml platform for on-device continual learning with quantized latent replays. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 11(4), 789-802.
- [39] Lin, J., Zhu, L., Chen, W. M., Wang, W. C., Gan, C., & Han, S. (2022). On-device training under 256kb memory. *Advances in Neural Information Processing Systems*, 35, 22941-22954.
- [40] G. Bellocchi, A. Capotondi, F. Conti, and A. Marongiu, "A RISC-V based FPGA Overlay to Simplify Embedded Accelerator Deployment", in *2021 24th Euromicro Conference on Digital System Design (DSD)*, 2021, pp. 9–17.
- [41] A. Bernardi, G. Brilli, A. Capotondi, A. Marongiu, and P. Burgio, "An FPGA Overlay for Efficient Real-Time Localization in 1/10th Scale Autonomous Vehicles", in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2022, pp. 915–920.
- [42] S. Machetti, P. D. Schiavone, T. C. Muller, M. Peon-Quiros, and D. Atienza, "X-heep: An open-source, configurable and extendible risc-v microcontroller for the exploration of ultra-low-power edge accelerators," 2024. [Online]. Available: <https://arxiv.org/abs/2401>
- [43] Enebuse, I., Ibrahim, B. K. K., Foo, M., Matharu, R. S., & Ahmed, H. (2022). Accuracy evaluation of hand-eye calibration techniques for vision-guided robots. *Plos one*, 17(10), e0273261.

- [44] Abdullah, M. W., Roth, H., Weyrich, M., & Wahrburg, J. (2015). An approach for peg-in-hole assembling using intuitive search algorithm based on human behavior and carried by sensors guided industrial robot. *IFAc-PapersOnLine*, 48(3), 1476-1481.
- [45] Suzuki, K., Yokota, Y., Kanazawa, Y., & Takebayashi, T. (2020, January). Online self-supervised learning for object picking: detecting optimum grasping position using a metric learning approach. In *2020 IEEE/SICE International Symposium on System Integration (SII)* (pp. 205-212). IEEE.
- [46] Ng, E., Liu, Z. and Kennedy, M., 2023, May. It takes two: Learning to plan for human-robot cooperative carrying. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 7526-7532). IEEE.
- [47] Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R. and Song, S., 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, p.02783649241273668.
- [48] Thomas L. Lee, Amos Storkey (2024) [Chunking: Continual Learning is not just about Distribution Shift](#). Third Conference on Lifelong Learning Agents (CoLLAs 2024)
- [49] Perry Gibson, José Cano, Elliot J. Crowley, Amos Storkey, Michael O'Boyle (2024) [DLAS: A Conceptual Model for Across-Stack Deep Learning Acceleration](#). *ACM Transactions on Architecture and Code Optimization*
- [50] Linus Ericsson, Miguel Espinosa, Chenhongyi Yang, Antreas Antoniou, Amos Storkey, Shay B. Cohen, Steven McDonagh, Elliot J. Crowley (2024) [einspace: Searching for Neural Architectures from Fundamental Operations](#). *Advances in Neural Information Processing Systems (NeurIPS)*
- [51] Thomas L. Lee, Amos Storkey (2024) [Approximate Bayesian Class-Conditional Models under Continuous Representation Shift](#). *International Conference on Artificial Intelligence and Statistics (AISTATS 2024)*
- [52] Velasco-Mata, A., Ruiz-Santaquiteria, J., Vallez, N., Deniz, O.. Using human pose information for handgun detection. *Neural Comput & Applic* 33, 17273–17286 (2021).
<https://doi.org/10.1007/s00521-021-06317-8>
- [53] N. Vallez, R. Rodriguez-Bobada, A. Dunne and J. L. Espinosa-Aranda, "Efficient In-Orbit CNN Updates," 2023 European Data Handling & Data Processing Conference (EDHPC), Juan Les Pins, France, 2023, pp. 1-5, doi: 10.23919/EDHPC59100.2023.10395959.
- [54] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, Yuntao Du. Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey. Feb 08 2024cs.CV, cs.LG
- [55] Ye-Xin Lu, Yang Ai, Zhen-Hua Ling, MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra, *Electrical Engineering and Systems Science > Audio and Speech Processing*, 2023



daiedge.eu