EPoSS
European Technology Platform
on Smart Systems Integration

AI AT THE EDGE

2021 – WHITE PAPER

Content

1 Introduction to "AI at the Edge" for Smart Systems

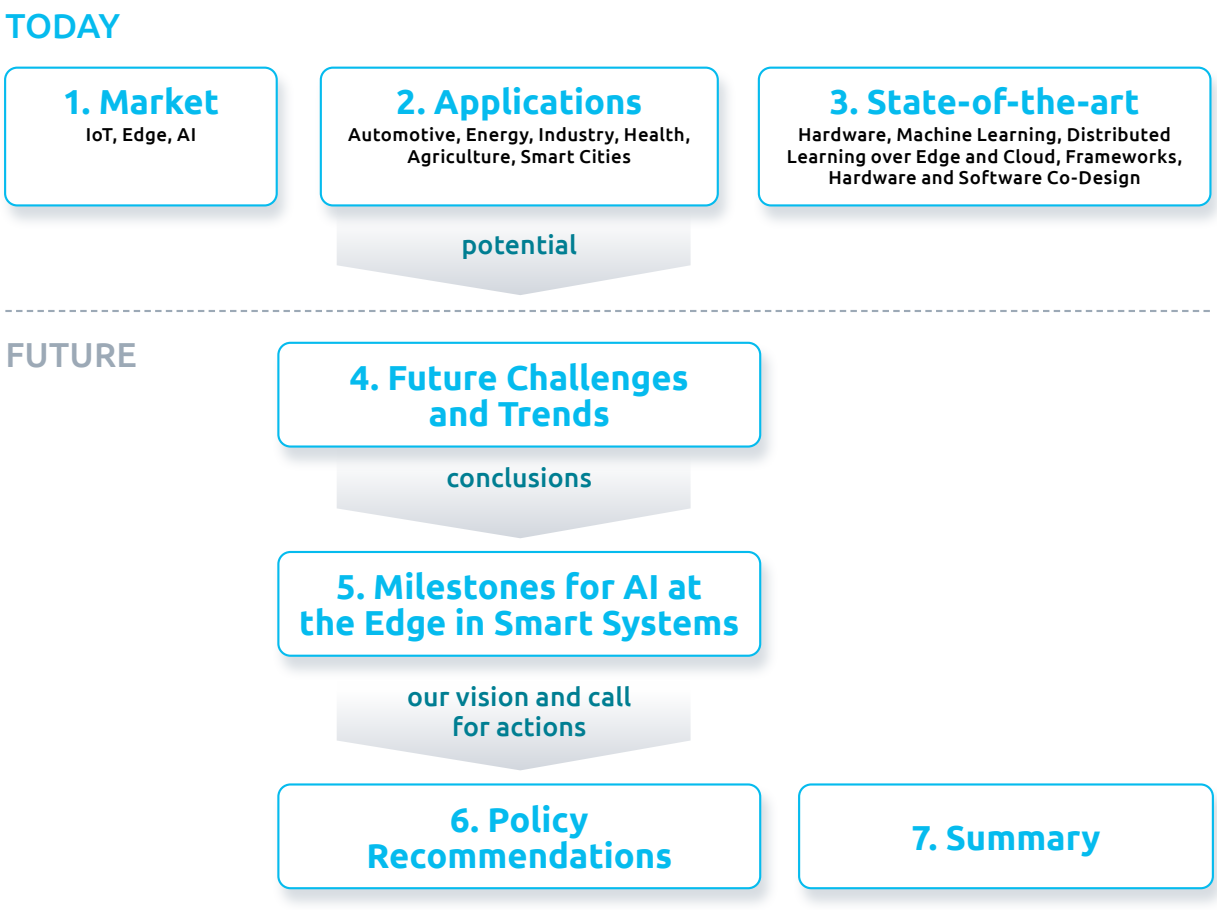
In this paper members of the European Platform on Smart Systems Integration (EPoSS) have collected their views on the benefits of incorporating Artificial Intelligence in future Smart devices and defined the actions required to achieve this to implement "AI at the Edge".

1.1 How to read this document

After a brief introduction to the topic and potential markets, in *Chapter 2* the authors describe the opportunities that lie in applications of AI at the Edge based on typical use cases and current R&D&I projects in Europe. The current state of the art is covered in *Chapter 3*. Using future requirements the cross-domain technological challenges for the next 10 years are summarised in *Chapter 4*.

To address individual needs of our audience, we divided the structure of this document into two parts: the first part focuses on status quo: *Chapter 1* includes the market potential of AI at the edge, *Chapter 2* describes the possible application domains and challenges and *Chapter 3* presents the *state-of-the-art* technologies that are available now.

The second part addresses the future. *Chapters 4* and *5* include the novel technologies, trends, and technological milestones that will drive the future research activities in the next ten years. *Chapter 6* outlines the recommendations of the experts to the political decision makes, in order to seize the full potential of AI at the Edge. Finally, the whitepaper concludes with the *Summary* of the major insights and recommendations.



1.2 Artificial intelligence at the Edge: introduction to the topic

1.2.1 DEFINITION OF "AI AT THE EDGE"

Artificial Intelligence (AI) is a technical system which has the ability to mimic human intelligence as characterized by behaviours such as sensing, learning, understanding, decision-making, and acting. Owing to the availability of powerful computing hardware (GPUs and specialist architectures) and of large amounts of data, AI solutions especially Machine Learning (ML) and more specifically Deep Learning (DL) have found numerous and widespread applications over the past two decades (such as image recognition, fault detection or automated driving functions).

Due to their reliance on large amounts of data, most current AI solutions require large-scale cloud data centres for computationally demanding processing tasks. Nevertheless, we are now in a new information-centric era in which computing is becoming pervasive and ubiquitous, thanks to the billion IoT devices connected to the Internet, and increasing digitalisation generates Zettabytes of data every year. Consequently, edge computing is emerging as a strong alternative to traditional cloud computing, enabling new types of applications (such as connected health, autonomous driving, Industry 4.0) with the advantage of implementing the required AI solutions as close as possible to the end-users and the data sources.

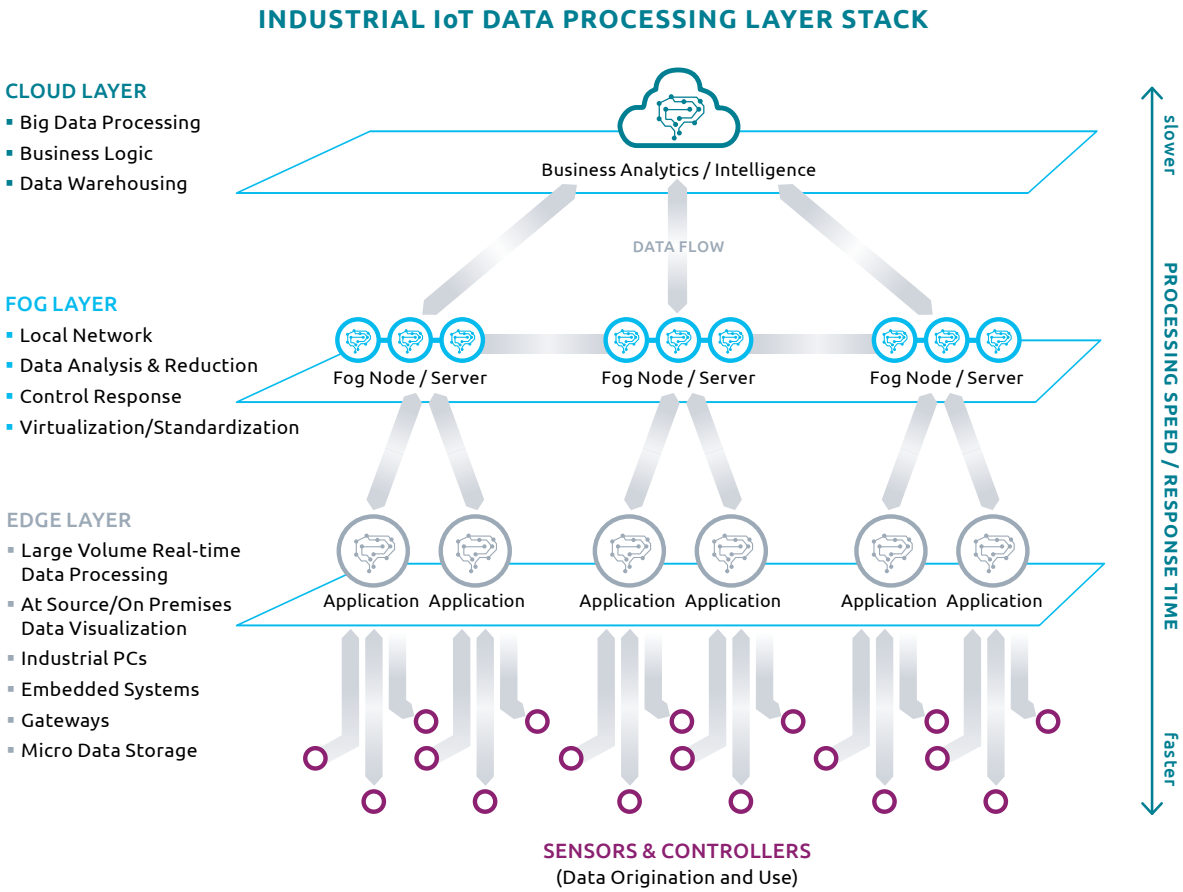


Figure 1: *Positioning of edge/extreme edge. The data processing stack for the Internet of Things consists of three layers: the edge layer, the fog layer and the cloud layer. In this paper we mainly address AI implementation at the Edge, close to smart sensors. The lowest layer represents the current use of AI-enhanced systems, often acting in a single system. As complexity and functionality increases, interaction between several AI systems is needed (between sensors for sensor fusion, between the AI model and a simulation model (Digital Twin) in Hybrid AI, or between a several AI systems). Ultimately, the highest complexity and functionality is achieved with distributed systems of systems (swarm AI, general intelligence).*

Although a consensus across academia and industry on a worldwide AI roadmap is still to be reached, one fact is that the pure dominance of cloud computing will come to an end. To exploit the full potential of AI, data processing solutions will run on distributed edge computing nodes, interconnected by next-generation IoT platforms and communications. In future functionality, energy consumption, stability, resilience, robustness and safety constraints will define their features. *Figure 1* shows an example of the interaction between cloud and edge computing as envisioned for the future.

1.3 Impact for the European Smart Systems industry and market

Factors driving the demand on edge computing solutions include a growing adoption of the Internet of Things (IoT) from 7% in 2019 to 12% by 2025 across industry[1]. Low-latency processing and real-time, automated decision-making and a need for processing exponentially increasing data volumes and network traffic require novel edge-based approaches. Moreover, the emergence of autonomous vehicles, wearable devices, connected infrastructures and the need for lightweight frameworks and systems (to enhance the efficiency of edge computing solutions) will create additional market opportunities for edge computing vendors. The Linux Foundation estimates in its "The 2021 State of the Edge"-Report[2] that between 2019 and 2028 up to $800 billion USD infrastructure investments will be rquired to cover the growing device and infrastructure edge demand. The expected investments into IoT devices and infrastructure edges will be relatively evenly split.

Technologies for wireless connectivity such as 5G are acting as a catalyst for market growth up to 35% CAGR alone for industrial IoT-solutions[3] with the total market for intelligent industrial edge computing (hardware, software, services) growing from $11.6B in 2019 to $30.8B by 2025.

Current leaders in cloud technologies see this as an opportunity to increase their market share and have started investing in the edge ecosystem by engaging in partnerships with global telecom companies and smaller innovative vendors[4]. It is quite evident that 5G, and its predicted benefits, has the potential to create a powerful network-based technology that is expected to reorganize industrial value chains[5].

Yole Développement forecasts for AI computing in consumer applications (in particular stand-alone and embedded sound and vision processors) a market increase from USD 2.3 billion in 2018 to 15.6 billion in 2024 at an average CAGR of 37.5%[6]. For AI in the automotive field (robotic vehicles, infotainment and ADAS) revenues starting from USD 174 million in 2018 to 13.8 billion in 2028 (average CAGR of 49%)[7], in AI for medical imaging from USD 332 million in 2019 to 2,886 billion in 2025 (CAGR of 36%)[8], and for neuromorphic computing and sensing an increase of the markets in the mobile, consumer, computing, automotive, medical and industrial fields from USD 112 million in 2024 to 25.993 billion in 2034 (CAGR of 64%)[9].

1.4 Technical advantages and opportunities of AI at the Edge

AI solutions that run autonomously, are distributed and implemented **at the Edge** offer the following advantages:

- **Increased real-time performance (low-latency):** Edge applications process data and generate results locally on the sensing device. As a consequence, the device is not required to be continuously connected with a cloud data-centre. As it can process data and take decisions independently, there is increased real-time performance in the decision-making process, reduced delay of data transmissions and improved response speed.
- **Reliable low-bandwidth communication**: Distributed devices can handle a large number of computational tasks, therefore reducing the need to send data to the cloud for storage and further processing. Overall, this results in minimizing the traffic load in the network and supports low-bandwidth communication.

- **Enhanced power-efficiency:** As the amount and rate of data exchange with the cloud is minimized, the power consumption of the device is reduced thus improving battery lifetime, which is critical for many edge devices.
- **Improved data security and privacy:** By processing data locally it does not have to be sent over a network to remote servers for processing. This improves data security and privacy as the data is not visible externally.

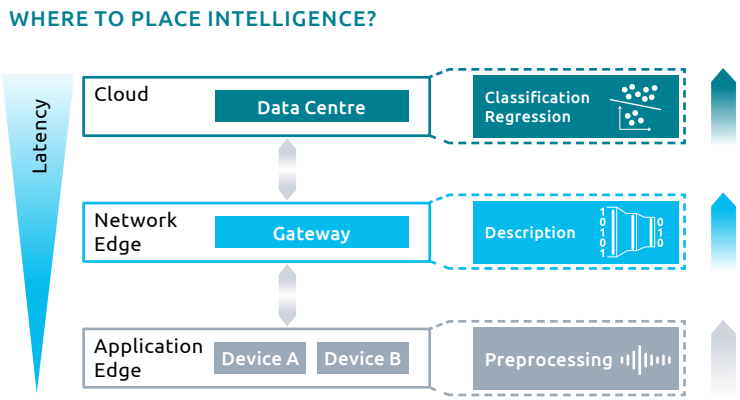


Figure 2: *Especially for applications that need real-time performance (low latency) the processing or pre-processing of data athe the Application edge is mandatory.*

1.5 Cost effectiveness

Over the past decades, the emergence and growth of the smartphone market and the mass production of semiconductors has allowed a significant cost reduction per unit for high performance processors that include AI capability. However, the required cost of an IoT device may limit the ability to include hardware components (memory, logic or storage) required for edge AI and this may limit the use of edge AI to high-end IoT products. Indeed, the applications currently driving the development of edge AI hardware are either computing-intensive (e.g. image processing for autonomous cars) or characterized by a high-level of criticality (e.g. organ-on-chip for health care).

IoT devices represent a wide range of connected objects, some will benefit from increased computing capacity but for low-power, remote and less data intensive devices the integration of high-end components to process AI may not be the best solution.

Sensor and Smart Systems solutions have evolved in the recent years, reducing the gap between the semiconductor chips and the final user/application. This strength allows a more tailored approach to AI system design that can mitigate the impact of components prices. This can provide adapted solutions for a wider range of devices and applications without a significant increase of cost when compared to “non-AI” solutions.

2 Applications for "AI at the Edge"

In this chapter, a number of applications involving the adoption of Edge AI solutions are illustrated and analysed with the goal to highlight the considerable breadth of scenarios where this technology can play an important role.

2.1 Automotive and multi-modal mobility

2.1.1 VEHICLE INTELLIGENCE AND V2X COMMUNICATION

Vehicle intelligence: While some advanced driver-assistance systems (ADAS), such as a lane keeping assistant or cruise control, are already commercially available. For several additional vehicle automation functions sufficiently efficient and reliable performance must still be developed and implemented before human drivers can be replaced by AI (in all operating domains). Transferring driving tasks successively from human to AI drivers and meeting all requirements with respect to sensing (scene understanding), decision-making and acting, presents a complex technological challenge with respect to both AI hardware and software models. Today it is clear that besides AI, the connectivity vehicle-to-vehicle (V2V) and between vehicles and infrastructure (V2I) will be key to deploying automated vehicles, since it provides the basis for the coordination of vehicles.

AI potential at the edge: Advances toward automated and ultimately autonomous mobility depend on progress in sensor and actuator technology, but most importantly on progress in AI technology. Each vehicle represents an edge node within the mobility system, connected to the cloud for services such as traffic or fleet management or mapping. Transferring AI tasks to the edge offers multiple benefits including improved system performance due to reduced communication and thereby processing latency, enhanced privacy or new functions such as driver authentication. The combination of vehicle intelligence and intelligent infrastructure using, for example, Multi-access Edge Computing (MEC) can provide further significant safety improvements[10].

Challenges: The optimal distribution of intelligence between the edge nodes (cars), the fog computing layer (e.g. traffic lights at an intersection) and the cloud (e.g. traffic management centres) presents a key, and strongly debated, topic in the field of vehicle automation. The answer is likely to differ for different operational domains, as automated shuttles on dedicated lanes require far less coordination from a central intelligence than an automated vehicle moving through dense mixed traffic (including non-automated, partially automated and fully automated vehicles).

2.1.2 OCCUPANT ACTIVITY "UNDERSTANDING"

Automated driving: Automated driving is one of the four main automotive trends[11], driven by technical developments, market expectations and continual legislative tightening. The technical solutions are focusing on the human-centred component, which covers two challenges: the Human-Machine Interface (HMI) and human perception of automated driving. Advanced HMI is the essential interface for seamless operation between the (semi)automatic system and humans. The EU-funded **HADRIAN** project[12] is developing a holistic driving solution, focusing on the utility of dynamically adjusting (fluid) human-machine interfaces taking environmental and driver conditions into account. On the other hand, human perception of driving style and safety is crucial for the acceptance of new technology through the increase of trust. The EU-funded **TEACHING** project[13] explores AI techniques at the Edge to realise the human-centred vision leveraging the physiological, emotional and cognitive state of vehicle occupants for the adaptation and optimisation of the autonomous driving applications.

AI potential at the edge: Managing transitions between different levels of autonomy is fundamental. The AI-based observer is a key point of this system as it detects the behaviour and the mental state of the driver. Edge AI offers the local calculation of the driver states, thus allowing for control of the response time thus

preventing personal data from leaving the vehicle and continuous learning to adapt the AI-based observer to each driver and passenger.

Challenges: Understanding vehicle occupants' physiological state and their ability to take over control of the (semi)autonomous vehicle are crucial for driving safety. Those challenges are further complemented by the need for the most effective interfacing to the driver. Those calculations must be performed at the local level to avoid basic connectivity risks. The inherent conflict between safety and AI is an open challenge, which is also complemented by the need for continuous learning.

2.1.3 AIR PATH CONTROL AND DIAGNOSTICS

Emission control and overall efficiency of the engine: The air path of an internal combustion engine is a crucial component for emission control and overall efficiency of the engine. The goal of air path diagnostics is to detect faults or poor performance and to identify the root cause.

AI potential at the edge: AI helps in efficiently executing the control strategy and in diagnostics of the air-path. For example, the heavyweight processing (e.g. physics-based simulations) used in executing the control strategy can be substituted with ML workloads. Compared to the original simulation model, the execution of the trained model implementation is less demanding. The deployed model can thus be executed on the edge. As sensors are mounted on the vehicle, this requires a split of intelligence between the backend (e.g. crowd-sourcing of vehicle-data to obtain the diagnostic model) and the edge (e.g. the various privacy related aspects).

Challenges: Challenging requirements such as time-predictability, dependability, energy-efficiency, and security need to be fulfilled. In this respect, the aim of ECSEL Joint Undertaking **FRACTAL**[14] is to create a reliable computing platform node, implementing a so-called Cognitive Edge with industry standards. This computing platform node will be the building block of scalable decentralized Internet of Things (ranging from Smart Low-Energy Computing Systems to High- Performance Computing Edge Nodes).

2.1.4 BATTERY LIFECYCLE MANAGEMENT

Predictive maintenance for battery aging: The in-use phase of a vehicle (road profile, climatic condition, driving, parking, charging) has a significant impact on battery aging. Batteries pose the risk of exploding ("thermal runaway") in normal use and the existing methods such as strain-, acoustic – and/or temperature sensors to detect thermal runaways.

AI potential at the edge: To better understand the aging behaviour of batteries, data-driven models based on aging experiments enable lifetime simulation and prediction. Predictive algorithms on the edge can crowd source data from vehicles and/or the lab. This provides critical correlations with battery safety and offers the potential of increasing the warning period. Within the **ECSEL JU Integrated Development 4.0**[15] a digital twin that allows the prediction of state of charge (SoC), state of health (SoH) and/or remaining lifetime is developed.

Challenges: There are a large variety of modelling approaches ranging from models using first principles (e.g. electro-chemical models) to purely data-driven models (needing to collect aging-related data in the lab and while operating the vehicle). These can be applied to the cell, module as well as package-level. Hybrid models have to be developed that aim to combine the models from first principles and data-driven models.

2.1.5 THERMAL MANAGEMENT OF THE POWERTRAIN

Energy control in the vehicle: The perception of the environment is carried out via vehicle and powertrain sensors coupled with the weather data and the traffic information that can be retrieved from dedicated service providers. The computing workload is split between processing in the backend (e.g. crowd-sourced data) and dedicated control units (energy control units) in the vehicle.

AI potential at the edge: AI has a tremendous potential in optimizing the energy efficiency based on the perceived environmental conditions. For example, by considering weather and traffic conditions to accelerate or delay the cooling process, AI-based strategies can augment classical model-based thermal control strategies. In line with such efforts, the Horizon Europe draft work programme 2021/2022, Cluster 5, mentions safe, seamless, smart, inclusive, resilient, climate neutral and sustainable mobility systems in terms of their expected impacts.

Challenges: The perception of the environment of the vehicle is key to optimize the vehicle's energy efficiency. Including the powertrain system (internal combustion engine, electric motor, fuel cell), energy storage system (hydrogen, electric), the passenger or cargo air conditioning system, and the traffic information (V2V, V2X). The collection of data offers the potential to characterize the context under which to perform the optimization. However, continuously collecting such data requires highly reliable connectivity (V2V, V2X) and an agreement on common mobility data sharing space[16]. In addition to standardization of data interoperability this includes data-lifecycle management that is designed around B2B, B2C and B2G data sharing.

2.1.6 AUTONOMOUS (INLAND) SHIPS BY PROJECT A-SWARM

Autonomous inland ships: could play an important role for the transportation of goods in big cities in the future. The German-funded **A-Swarm**[17] project explores this topic. It is planned in the project to fit a barge with near-field and far-field sensors as well as edge platforms with AI accelerators.

AI potential at the edge: The accelerators will be used to run AI models that locally process the data from sensors and control the engines of the barge. Allowing for the implementation of a system with the necessary real time capabilities required to traverse the waterways in cities.

Challenges: To realize this system many different problems need to be solved. One of the biggest is to filter out, in real time, the noise and movement generated by the water from the sensor data. Furthermore, it needs to be explored how different intelligent edge systems can efficiently communicate with each other to enable for example an automated unloading of the barge.

2.1.7 MASSIVE SENSOR TECHNOLOGY AND NETWORKS

Massive sensor technology and networks: The new concepts for autonomous mobility, digital industry, and decentralized bidirectional and multi-modal energy supply, as well as smart city and smart home applications, require massively more sensor technology and electronics with significantly higher performance in each individual product than today. At the same time, reliability and safety requirements are increasing dramatically, as the operation of automated and autonomous systems are no longer overseen by human operators. Instead, the lives of passengers, the economics of production, and the stability of utilities depend entirely on the functionality of their electronics. The current way to ensure the highest standards of safety and availability relies mainly on redundancy at all levels of integration (including full system redundancy). This is very expensive, resource heavy and sub optimal. The failures can occur without warning and in both the primary and the redundant unit. Therefore, the ultimate fallback solutions have to be used quite often (e.g. emergency stop). They are safe but usually mean the sudden end of operation. This approach would lead to an unreasonably low availability of ultra-complex systems like autonomous cars. New strategies with smart and pro-active safety assurance need to be developed that are based on continuous self-monitoring, remaining life estimation, and active failure prevention in the electronic systems.

AI potential at the edge: An intelligent approach to functional safety will achieve a higher level of confidence and trustworthiness with less redundancy than today. It will require the inclusion of artificial intelligence algorithms as essential elements. Trained by data from comprehensive physics of failure (PoF) studies and big data-driven (DD) analyses, compact AI routines can be developed and implemented directly into the individual products to deliver maximum availability.

Challenges: Despite the limited computational resources at this edge position, the AI routines should cover the current application scenario very well. This can be achieved by developing dedicated meta-models and by resource-optimized programming. In addition, a (non-permanent) connection to the cloud server allows dynamic updates to best adapt to changing scenarios (e.g. from summer to winter conditions) and to continuously improve the meta-models (e.g. by learning from the entire fleet). A number of projects have already started to explore this approach to AI-based smart safety solution for electronic systems, e.g. ECSEL **iRel4.0**[18] – PoF and DD analyses, ITEA3 **COMPAS**[19] – compact models for AI, H2020-GV **EVC1000**[20] – early warning indicators, lifetime estimation. However, the main part of the research work in this area is still ahead.

2.2 Energy

2.2.1 SMART GRID

Distributed energy sources: The development of Smart Grids over the past two decades was a necessary response to the fundamental shift from a unidirectional supply of electricity (from power plants to consumers) toward an increasingly decentralized, bidirectional and complex network. The widespread use of renewable energy sources has resulted in a corresponding growth in the number of network nodes. Advances in ICT have enabled smart home applications which, alongside the introduction of electric vehicles, constitute new agents and further increase the complexity at individual network nodes. At the same time, the introduction of smart meters at network endpoints and ubiquitous sensors throughout the grid, have added a digital layer comprising a myriad of sensors and providing large amounts of data. This data availability and the increasing diversification and distribution of energy sources and applications call for an equivalent distribution of intelligence throughout the grid, to maximise network efficiency, optimize grid management and enable new (end-user) applications, including data privacy.

AI potential at the edge: Large amounts of data concerning energy demand and supply accumulate at individual network nodes and must be processed efficiently at the Edge to exploit their full value. Machine learning applications for Smart Grids include classification and clustering models for big data processing, are used primarily by utility suppliers and cloud service providers to group consumers according to their usage patterns and apply predictive models for future demand[21]. Prediction models can be used for the supply of renewable energy when weather forecasts are included. While many of these models can be applied for management, decision-making and control processes at the (micro) grid level can be run in cloud data centres or fog gateways, some applications potential optimisations can only be fully unlocked using edge AI. Cognitive applications of edge computing in Smart Grids include intelligent agents used both for energy market issues (management, pricing and scheduling) and for network management (security, reliability, fault handling and efficiency). Possible use cases include:

- Combination of AI and blockchain technology for the integration of electric vehicles in power management platforms for Smart Grids[22].
- Dynamic pricing to balance demand and supply[23].
- Pre-processing strategy of hierarchical decision-making to optimise resource usage based on service level requirements.

- Data-driven methods to analyse equipment and end-user behaviour in the distribution network, for example, to provide energy as a service (EaaS)[24].
- Fault detection and diagnosis in the transmission grid (e.g. video surveillance and scene interpretation using drones).
- Real-time monitoring[25].

Challenges: A central challenge for the application of edge computing and AI in Smart Grids remains the design and implementation of efficient system architectures that meet the real time and safety requirements on AI at the device, network and application level and distribute tasks as well as required intelligence between cloud, fog and edge.

2.2.2 SMART BUILDINGS

Buildings as networked cyber-physical energy systems: Buildings are a major contributor to the overall energy consumption. Since passive means (e.g. thermal insulation) are nearly fully exploited, smart buildings are envisaged to be the future enabler for further improvement in energy efficiency[26]. The objectives of building energy control systems are multi-dimensional and complex aimed at using a minimum of energy (preferably generated on-site from renewable sources), a prescribed level of comfort and a healthy indoor climate must be provided. Since the components of the building energy systems are integrating more sensors and embedded systems, buildings are becoming networked cyber-physical energy systems – especially larger objects like airports, shopping malls or office buildings.

AI potential at the edge: A high number of multivariate sensors are required to exploit the full potential of model predictive control schemes besides standard parameters such as temperature, humidity, CO_2 and the occupation of rooms, which are usual inputs to the control system. While the main control system is usually implemented as a centralized controller, there are relevant applications for data analytics and AI on edge devices and smart sensors; for example: the number of persons present inside a room is a relevant input parameter for building controls. Image sensors allow for a precise counting of people. But to enable a required level of privacy, raw image data should not be spread among open data networks. Implementing AI algorithms directly on the device can help to analyse the image in order to extract the relevant information for the control system. Sensors in energy system components, like fans or air filters, are an enabler for predictive maintenance schemes, allowing higher efficiency and reduced maintenance costs. Using wireless technology, easy installation or retrofitting would be possible - especially at places that are hard to reach by the tethered data network. However wireless data transmission from basements can be difficult due to the metal structures in heating, ventilation and air conditioning systems. Data can be reliably transmitted, with a reduced bandwidth, by using data analytics at the sensor to provide only the relevant information on the current status of the component instead of time series data from pressure sensors etc.

Challenges: Only with a large number of sensors can explore the merits of energy savings, i.e. the economic benefit per sensor is rather low. In turn, a smart sensor for a building energy system must be a rugged and a low-cost system. Furthermore, in many use cases, wireless connectivity is strongly demanded. An optimized power consumption ensures long maintenance intervals, imposing challenges on energy efficiency of the onboard data acquisition and processing.

2.3 Digital Industry

2.3.1 PREDICTIVE MAINTENANCE

Industry 4.0 and predictive maintenance: Industrial applications of IoT are predicted to generate a significant economic benefit. Predictive maintenance is a popular example enabled mainly by the analysis of huge amounts of data generated by sensors integrated into industrial assets. This is implemented by the classification of the acquired data, with respect to the status of critical components, and using prediction models to enable a forecast of remaining lifetime and, in turn, to optimize maintenance schedules.

AI potential at the edge: Implementing AI on the edge devices near the sensors would offer several benefits: reduction in the transmitted data volume, which is particularly important for sensors generating large data streams such as vibration time series. Data from heterogeneous sensors can be fused on the device. This also enables cross-validation of sensor data, improving the resilience of the system. Local data analysis can reduce the latency of the AI compared to a cloud based solution, this can be an important advantage when detecting critical faults.

Challenges: In order to gain economic benefits from the sensor signal analysis, the accuracy of the algorithms has to be very high. False alarms or undetected failures can cause severe financial losses or even damages to equipment. Another important aspect is the availability of training and validation data. Only for mass production lines, the necessary amount of representative data can be collected in a reasonable time. In cases of more individualized production, algorithms have to cope with small training sets; or the application of synthetic data from simulation model scan be considered.

2.3.2 RELIABLE PREVENTION OF EARLY FAILURES

Reliable prevention of early product failures: Product reliability has a typical characteristic. A relatively high failure rate occurs during the first operating period. These early failures are caused by the small variations in material, shape, or process properties during fabrication. None of these stochastic deviations exceeds their specified limits, so current process control algorithms cannot detect the reliability risk that arises from unfortunate combinations of these variations.

AI potential at the edge: Expanding the scope of process control, by including a larger number of process steps in advanced data analysis using artificial intelligence schemes, can detect a significant portion of these risky combinations of inherently permissible variations. The ECSEL project **iRel4.0**[27] explores this approach with the example of microelectronic production. While the core part of AI-based data analysis can be performed by the large computer clusters, that provide general process control at the manufacturing site, additional edge capabilities are required to enable corrective countermeasures to be taken in real-time at all relevant process tools to provide the important data in a pre-aggregated form.

Challenges: The computational edge capabilities are thus an essential part of the overall AI system. The flexibility, latency, and security requirements of advanced process control cannot be met without them.

2.3.3 ROBOT CO-WORKING

Collaborative robots in industrial environments support human workforce in the fulfilment of repetitive jobs or heavy lifting, for instance. Applications can be found mainly in the manufacturing industry, e.g. assembly of automotive parts.

AI potential on the edge: Edge AI enables new possibilities for the cooperation of humans and robots, because in contrast to cloud based systems edge AI is fast enough to handle situations where the robot could inflict harm. To implement these new possibilities sensors need to be deployed that are able to monitor the

environment and the movement of humans and animals within range. The data from these sensors are locally processed by AI models in the robot or running on nearby edge nodes. Afterwards, edge AI-based components use the processed data to control the robot allowing for close cooperation with humans in performing complex tasks like the manufacturing of custom products in workshops or rescue operations. To implement this vision of close cooperation many challenges need to be solved such as training the robots for new tasks.

Challenges: In addition to more traditional robotic applications, the safety of the human worker has to be considered, since the robot and the human share a common working space. Operational strategies ensuring safety of the worker require advanced sensing capabilities of the robot[28]. In addition, the sensor data, e.g. from an image sensor, has to be processed with low latency in order to enable a quick reaction of the robot in a critical situation. Thus, transferring cognitive and analytic capabilities to the edge, i.e. a single robot, is advantageous. Potential strategies include distribution of AI methods in a network of robotic devices[29].

Finally, reliability and functional safety requirements of the robotic system with integrated AI capabilities have to be met during the design process.

2.4 Health and wellbeing

2.4.1 VITAL SENSING WITH RADAR

Vital sign monitoring based on radar sensors: will be an important component in many medical applications. However, a cloud-based implementation of the sensing would be too slow in time critical contexts. This is not the only problem of cloud systems as storing generated data in them is also a privacy concern.

AI potential at the edge: Issues of latency and privacy can be solved by using edge AI. When the radar data is processed locally, the information about heart rate, respiration and so on are available fast enough to trigger other parts of the system that can save the life of the human. Furthermore, the results can then be deleted or anonymized before they are sent to the cloud.

Challenges: The accuracy of current edge AI implementations of such products is too low to avoid high false alarms rates. Hence, the accuracy of algorithms needs to be improved to enable better adoption of life saving applications.

2.4.2 PERSONALISED MEDICINE

Personalised Medicine: Human physiology can vary greatly from individual to individual. Examples for that include blood pressure or lung capacity. However, these differences need to be considered for accurate medical applications like vital sign monitoring. Due to privacy concerns, it is difficult to process this information in the cloud-based solutions.

AI potential at the edge: Edge AI offers the possibility of maintaining privacy when processing medical data. Furthermore, many medical applications require real time processing, which can be better realized with local AI. By exploiting these two aspects, many medical and consumer applications can be implemented which were not possible in the past. For example, different organisations work on integrating sensors and AI into clothes allowing for feedback loop based training of athletes.

Challenges: Processing data at the Edge does not make it totally safe against malicious access. Hence, the security measures of edge AI processing pipelines need to be further improved to ensure that medical data or applications are not misused.

2.4.3 AFFECTIVE COMPUTING ("AI OF EMOTIONS")

Detecting and measuring human emotions: Affective computing is interested in automatically detecting and recognizing the emotional state of a human either with remote or "nearable" sensors (visible and IR imagery, audio, physiology), or with sensors in contact (wearables) for physiology, or activity monitoring. Emotions, a classic conceptual representation of which follows a 2D valence (negative / positive) versus intensity (calm / excited) pattern, have an essential role in human behaviour. These influence the mechanisms of perception, attention, decision making, and social behaviour. The purpose of estimating emotional states is to improve understanding of human behaviour. This is the strongest reason as emotional states are both very personal and evolving, very different from one individual to another, and from one situation to another.

AI potential at the edge: The edge AI allows for maintaining the confidentiality of the data inside the measurement device, to guarantee the autonomy of the devices, and to aim for an individual estimator learning over time. The objective of the studies conducted at the CEA LETI is to develop an autonomous and ambulatory stress observer based on physiological signals, aimed at self-assessment and coaching for well-being (see **M.O.T.I.O.N** project)[30].

Challenges: Privacy and personalisation. On the road to individual guidance –whether medical or for other purpose (wellbeing, sports or emotion management) – local processing of data answers potential issue of confidentiality and data protection. In addition, the use of AI allows identification and adaptation to individual response pattern to the targeted monitoring (activity, treatment…). Once anonymised, this individual response (learned and characterised thanks to the AI) can feed wider models so that it can be shared and benefits to other users/patients and helps them in managing their own activities.

2.4.4 SPORT ANALYTICS

Prevalence of lower-limb injuries: Lower-limb injuries are common among athletes, accounting for 77% of hospitalized sport-related injuries, and are a risk factor for early-onset osteoarthritis. High-impact forces are one of the factors contributing to lower-limb injuries. To decrease the prevalence of lower-limb injuries, and their associated long-term disability and economic burden, multiple injury prevention programs have been proposed. These take into account the study of ground reaction forces (GRFs) in order to enhance athletes' performance, determine injury-related factors, and evaluate rehabilitation programs' outcomes.

AI potential at the edge: Together with industry partners, the Tyndall National Institute have developed a miniaturised monitoring system, integrating ultra-accurate accelerometers and neural networks, to estimate the impact GRF forces while running. Besides being a unique solution for multiple injury prevention, the developed solution can be used by elite athletes, sports teams, coaches, scientists, and consumers who would use novel performance monitoring systems to keep pushing the boundaries of their sports and gain performance advantages.

Challenges: Major challenges in the system implementation are related to the development of a neural network that is sufficiently accurate to model GRFs while it is also simple enough to be deployed on a resource-constrained microcontroller with limited energy consumption. Moreover, an open challenge is related to the deployment of personalized athlete-specific models rather than general-purpose neural networks; this could be achieved by either training a whole network from scratch directly on the wearable unit by relying only on the data collected from the individual athlete, or by adopting a transfer learning approach where a number of layers in the general-purpose network are trained based on the data from all the available subjects and are frozen and deployed on the microcontroller and the data collected from the individual athlete are used to train only the last layers of the deployed neural network.

2.4.5 PHYSIOLOGICAL MONITORING

Health markers: Elevated blood pressure is a major health concern and a risk factor for complicated cardiovascular morbidities including coronary heart disease, ischemic, and haemorrhagic stroke. WHO reported an estimated 7.5 million deaths due to elevated blood pressure. The accurate measurement of blood pressure is important to timely detect health threats. Therefore, to get a continuous, accurate, and reliable insight into a person's cardiovascular health condition requires a practical approach.

Sepsis is also another good example, one of the leading causes of death worldwide, with incidence and mortality rates failing to decrease substantially over the last few decades.

AI potential at the edge: Edge computing is experiencing a massive growth in healthcare applications as it helps to maintain the privacy of patients (e.g. data is locally processed, without engaging cloud services in the overall process), and allows a fast and real-time decision support system.

One of the objectives of the **HOLISTICS** project led by Tyndall National Institute[31], in cooperation with its industry partners, is the adoption of edge analytics into wearable devices for health-related use case scenarios (e.g. blood pressure monitoring). Cuffless blood pressure monitoring devices adopting AI solutions based on the analysis of PPG or PTT signals have shown promising results in recent years.

As an example to illustrate the value of edge-based AI models in the management of vital signs, the model can raise timely alerts pro-actively prompting clinicians without needing time-consuming and costly laboratory tests. AI solutions have, therefore, the potential to be used on wearable devices to predict the prognosis (e.g. blood pressure), and/or detect the pathogens causing an infectious process (i.e. sepsis).

Challenges: A typical challenge of health-related datasets is the presence of a high imbalance in the data. The development of the outcomes for patients with sepsis and recommend the treatment process (e.g. the medications to be used during sepsis), of techniques and approaches able to tackle this problem at a technical level (i.e. data augmentation, resampling techniques) and policy level (e.g. data collection process, data sharing policy, new standards) is diffusing steadily. Moreover, the possibility to provide tailored medical treatment (e.g. personalized medicine) is attracting increased attention over the recent years; however, its implementation and deployment into edge devices in real-world scenarios is still in its infancy.

2.5 Agriculture, Farming and Natural Resources

2.5.1 AUTOMATED WEEDING

Chemical weeding to reduce the competition between weeds and crops: Vegetable production imposes a wide variety of farming operations because of the diversity of crops and the related planting parameters such as the seedbed structure, the seeding density, the spacing between rows and the distance between plants in each row. In addition to these agricultural operations, vegetables require early weeding (7 to 15 days after sowing or planting) due to the strong competition between weeds and crop and the increasing difficulty of removing weeds without damaging the crop. Once the crops cover all the row, weeds are stifled as soon as they appear, and weeding becomes less critical. Chemical weeding is the classical solution to reduce the competition between weeds and crops. However, the growing consumers' demand for product quality and for the absence of phytosanitary residues, is having an increasing impact on agricultural practices. Mechanical weeding (hoeing) is, therefore, increasingly necessary. Nevertheless, it remains difficult to implement weeding within the rows, because destroying the weeds inside a row while preserving the plants is very delicate, especially when the sowing is dense. To date the only mechanized or automated solutions concern inter-row weeding (weeding between two rows).

AI potential at the edge: Commercial AI-based intra-row weeding solutions exist only for crops with significant inter-plant distances (lettuce or cabbage for instance). No automatic hoeing solution exist for carrot, peas, beans, sweet corn, onions, etc. To realize intra-weeding for these crops some AI-capabilities in the weeding machines are required to adapt to changing environment in real time.

Challenges: A stable connection to the cloud cannot be guaranteed on fields all the time. Furthermore, the automated weeding machines should be as power efficient as possible. Both of these requirements could be solved by neuromorphic AI algorithms, due to their lower energy demand compared to standard neural networks at the Edge. Such algorithms and corresponding hardware are explored in European funding project **Andante**[32] but these topics will require much more work than which can be achieved within one project.

2.5.2 DRONES FOR PRECISION AGRICULTURE

Precision agriculture is one of the scenarios where Unmanned Aerial Vehicles (UAVs) or drones are currently being used and demonstrated. They are equipped with cameras and sensors which allow taking close images of the crops, field operations and of the machines.

AI potential at the edge: This information can be used for tasks such as obtaining Normalised Difference Vegetation Index (NDVI) maps from multispectral cameras which can support decision making about spraying or perform additional operations in the crops, for example to recognise areas that may be affected by pests and to apply phytosanitary or pesticide treatments. They can even act as a network gateway to collect information from IoT sensors using low-cost and wide area network protocols like LoRaWAN (Long Range Wide Area Network).

Challenges: The deployment and the usage of drones and UAVs in the agriculture domain still presents challenges that must be solved, e.g.

- be intelligent enough to fly autonomously without requiring major interventions from specialised human operators
- be capable of dynamically readjusting the missions based on context information coming from onboard sensors and other sources of data deployed in the crops
- collaborate with other drones or ground robots to perform more complex tasks in complex and larger terrains
- guarantee compliance with security regulations and incorporate trustworthy requirements and guidelines.

To address most of the previous points, artificial intelligence processes will be embedded directly on drones and robots in order to increase their autonomy and real-time capabilities.

2.5.3 SOIL CONTROL

Efficient food production is important to ensure the food supply of mankind. Hence, more and more sensors are deployed around, and in fields to gather data about their state and planted crops. For soil monitoring Biodegradable sensors are being researched. The idea is to mix them into the fertilizer, which is then put on the field. Afterwards, they send their data for between six months and one year to a node near the field and this node transfers the data to the cloud.

AI potential at the edge: The amount of data generated by the field monitoring sensors is very high. However, not all of the data is relevant and can be averaged over multiple sensors e.g. the average soil moisture level of a field. Edge AI can be trained to analyse these large data volumes resulting in lower amounts of data needed to be sent to the cloud as well as lower network load.

Challenges: The critical point about bio-degradable sensors is that they cease to work after a specific amount of time. In addition, other sensors deployed around the field may have a lower average lifetime than sensors in other contexts. This requires the Edge AI solutions for this application to be able to handle fluctuating amounts of incoming data. Such high levels of flexibility are not well explored yet.

2.6 Smart Cities

2.6.1 SMART STREETLIGHTS

Lighting systems adjusting the brightness to the individual conditions of the surroundings: Smart streetlights could provide important services for smarter and greener cities in the future.

AI potential at the edge: Using different kinds of sensors and edge AI, the streetlights can detect whether, and at what speed, a pedestrian or motorist is approaching. As long as the person is within the radius of the light, this area is illuminated by built-in LED lamps. If the person moves away, the lighting is reduced. In adverse weather conditions, such as snow or rain, the light output could be increased automatically as required. The edge AI evaluating the sensor data can run on microcontrollers in the lamp or on other nodes in the proximity. This dynamic light regulation saves energy and costs. Smart streetlights could also be used for implementing other important services like the charging of electric vehicles and the measurement of the air quality.

Challenges: A central challenge of this application is managing the access to the results of the Edge AI. The processed sensor data can be of interest to different parties, for example for the police in case of accidents or insurance services that insure shops near the smart street lights. One approach to solve this challenge would be to combine block chain technologies with Edge AI. However, this is a research field which is still in its initial phase.

2.6.2 AIR QUALITY

Improving air quality using gas sensors: Gas sensors currently available on the market are often quite unstable, inaccurate and show large cross sensitivities to other interfering gases. Moreover, they are often very large (not in a portable form factor) and quite costly.

AI potential at the edge: Neural networks at the Edge are crucial to gas sensing especially when it comes to accurately identifying different gases in an outdoor environment. While the sensor technology itself (materials, geometry, temperature modulation, number of sensing fields, etc.) can surely help to improve sensitivity to target gases, algorithms play a very important role when it comes not only to classifying gases but also to quantifying them in parts per billion (ppb). Since gas sensors in most use cases have limited connection to the internet, these algorithms need to be deployed on the sensor node.

Challenges: Recent results already show that traditional neural networks can strike the right balance between accuracy and robustness for air quality monitoring, still many open questions remain on the behaviour of air quality monitoring sensor deployed in the field over a long time. Here, it is even more crucial to ensure long battery life and wider online learning at the Edge for specific use cases and more self-diagnostics on the performance of the sensor.

2.6.3 ALARM SYSTEMS

Increased safety with intelligent Alarm Systems: Edge AI-based alarm systems are a good example of how edge computing solutions enrich existing smart building systems.

AI potential at the edge: While previous alarm systems use a microphone and simple threshold rules to detect glass breakage when an unlawful entry is made into an apartment, the new generation of alarms process infor-

mation from multiple data sources via data fusion and neural networks. This minimizes the number of false alarm and significantly increases the reliability of the system. This solution can easily be integrated into existing glass breakage alarm systems.

Challenges: As these new generations of alarm system become more widely spread, they will become targets for attacks. Currently little work has been done in the area of securing neural networks, making them responsible for the whole system.

3 State-of-the-art of "AI at the Edge"

3.1 Edge AI for smarter systems

Edge AI resides at the location where the virtual world of the network hits the real world, where sensors and actuators are the link.

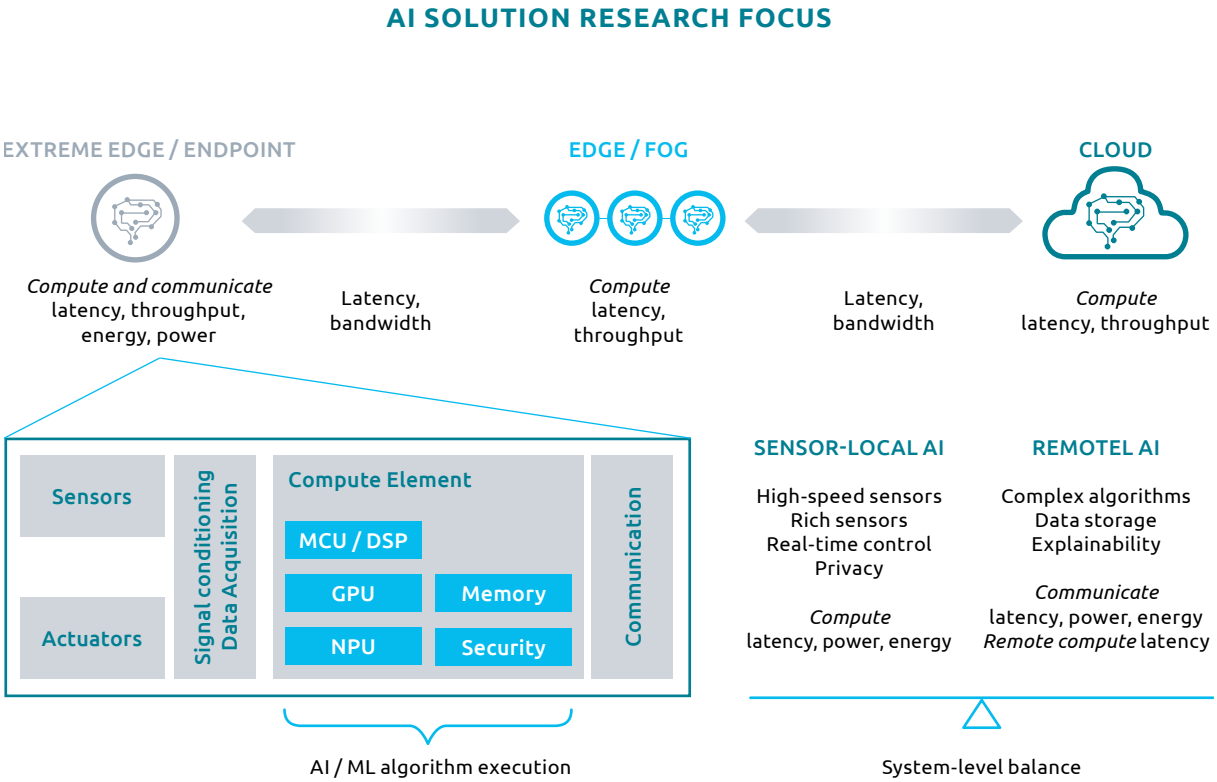


Figure 3: *At the system level, the place to run AI algorithms depends on multiple factors and is often a balancing act between the time and energy cost of local compute vs remote compute. Algorithms can be distributed at multiple levels as well. The balance point will shift over time, following advances in wireless technologies and neural processing.*

Edge computing can be mainly segmented in the following areas:

- Hardware (HW)
- Software
- IoT Platforms/Communication
- Services

Those areas are separately discussed below.

3.2 Hardware for edge AI

Choosing the ideal hardware for a particular application requires careful consideration of all the requirements. A successful system design finds a balance between the different aspects of system architecture, such as memory footprint, executing time, model accuracy, power consumption, scalability, cost, and maintainability. While data-centres allow engineers to scale available computational power to the current demand (via GPUs or TPUs), an application running on edge devices needs to keep sufficient power reserves. An increasing number of vendors are now moving from producing simple resource-scarce microcontrollers (e.g. ARM Cortex MCU) to pairing general-purpose processors with specialized units tailored to execute the computational tasks required to implement AI solution. As embedded systems are typically focused on using AI in the form of machine learning (ML) for interpreting incoming sensor data, these specialized sub-processors aim to speed up a classification or prediction tasks while maintaining a low power draw. This is especially important in applications running on battery power or with a low potential for cooling the system.

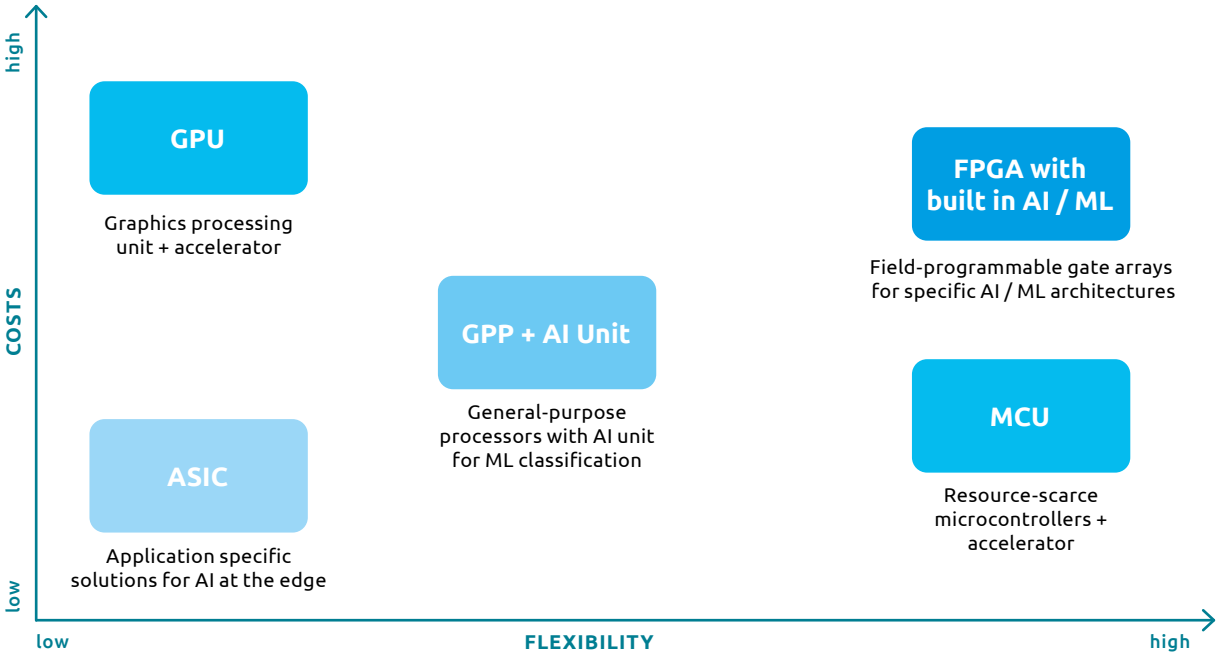


Figure 4: *Depending on the particular application requirements, different types of hardware are available and have to be chosen for Edge AI realisation.*

A common approach is the use of HW accelerators which can be either directly co-located with the general-purpose processor on the same silicon or might be connected as a separate chip. These accelerators stretch the power continuum from relatively simple digital signal processors (DSPs) to highly parallel matrix computation units and similar advanced designs. These accelerators can either be monolithic designs such as special Edge variants of Google's TPU or a distributed set of smaller compute cores. Some examples for the latter are the Tensor Cores in newer Nvidia GPU architectures, the Hexagon cores in Qualcomm SnapDragon SoCs and Intel Movidius SHAVE processors. They commonly need specialized drivers and software libraries that allow software developers to take advantage of their capabilities.

Currently, many accelerators rely on reduced precision computation, replacing costly floating-point mathematics with lightweight integer operations of 8bit precision, or even lower. Co-locating memory close to computation

cores and pre-loading repeatedly used variables cuts down on time spent shuffling data around. Nowadays some accelerator hardware architectures are closely tailored to the operation of specific ML algorithms, allowing for very efficient computation. This specialization however comes at the cost of flexibility in a rapidly evolving field.

A solution is presented using field-programmable gate arrays (FPGAs), which are becoming heterogeneous platforms that combine powerful CPU systems, specialized arrays of AI accelerator cores and traditional fabric, where the hardware can be programmed by the use of hardware description language (HDL) or C/C++ via high-level synthesis tools. FPGAs combine the benefits of specialized hardware with the freedom to change the layout even after the chip has left the factory. The use of FPGA can create more dynamic, scalable, and flexible systems, even though they often carry higher cost.

Another type of processor is emerging as a new class of processing accelerator for these predominantly data-centric heterogenous processing tasks as offload to the main CPU. These processors are called manycores and are referred as DPUs (Data Processing Unit) in the industry. For example, in the case of a car or a drone, the challenge is to integrate the AI in complex, heterogeneous, real-time systems especially regarding pre-processing, DL-based processing, and post-processing. Intensive mathematical algorithms, signal processing, network or storage software stacks in the context of end-to-end use case is critical to meet key requirements such as form factor/size, power consumption, and costs. DPU or manycores provide a solution for such complex requirements. One pioneer company in such processors is Kalray with its manycores MPPA (Massively Parallel Processor Array) solution.

Today, a system architect needs to weigh both current and future requirements of their systems when deciding on which combination of conventional and specialized computation cores to pick. Other than datacentres, where upgrades are done under controlled conditions, devices out in the field are harder to upgrade to more capable hardware, especially when faced with the number and variety of different customers.

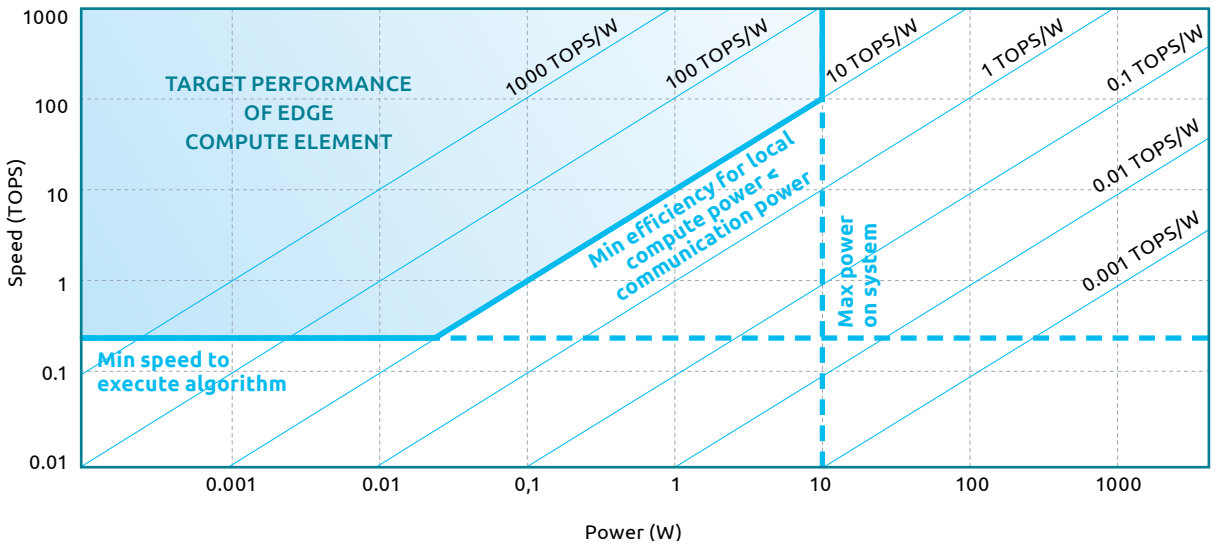


Figure 5: *The system analysis results in constraints for computation speed, power consumption and power-efficiency for the compute element, given a specific algorithm. The target performance zone is different for each application or application domains.*

For each application, one can balance the energy cost and latency between computing the AI locally or transferring the data to cloud for processing. Typical mobile applications are limited by power constraints and a maximum laten-

cy requirement. The performance and power efficiency of the local computation solution should be better than the performance and efficiency of the communication system. For instance, an AI processor on a small drone, tasked with running the tiny-YOLOv3 algorithm, should consume less than 10W, and provide more than 336 GOPS compute power able to run at 30 frames per seconds. Its power efficiency must be better than 112 GOPS/W (Giga Operation Per Second Per Watt) to be competitive with current 5G data transmission and computation at the edge. It is clear that improvements in communication level will push the minimal requirements for local AI computing even higher.

Even though vendors are typically focusing on ever more capable H/W accelerators, new development tools and libraries of algorithms and software can also contribute to boost the system performance.

3.3 Machine learning models for edge AI

Current AI models for the edge are far more limited in terms of performance when compared to cloud-based models because of the relatively limited computation and storage abilities. Model training and inference on resource-scarce devices are still a debated problem throughout academia and industry.

A number of novel libraries and algorithms have been developed in the recent years with the goal to adapt standard ML models to resource-constrained devices. A well-known example is given by ProtoNN which aims to adapt kNN in memory space-limited microcontrollers via sparse-projection and joint optimization. For low memory scenarios (< 2 kB), ProtoNN outperformed the state-of-the-art compressed models. In settings allowing 16-32 kB memory, it matched the performance of the state-of-the-art compressed models. Moreover, when compared to the best uncompressed models, ProtoNN was only 1–2% less accurate while consuming 1–2 orders of magnitude less memory.

Bonsai is another novel algorithm based, instead, on decision trees and aims to reduce the model size by learning a sparse, single shallow tree. When deployed on an Arduino Uno, Bonsai required only 70 bytes for a binary classification model and 500 bytes for a 62-class classification model. Its prediction accuracy was up to 30% higher than other resource-constrained models and even comparable with unconstrained models, with better prediction times and energy usage.

The development of neural networks and deep neural networks with lighter and faster architectures (e.g. small size model, minimization of trainable parameters, minimization of the number of computations) for edge platforms has also gained massive traction among researchers. Some examples are represented by CMSIS-NN (developed for Cortex-M processor cores) which generates neural networks that can achieve about a fourfold improvement in performance and energy efficiency, yet minimizing the memory footprint.

Even recurrent neural networks (RNN) have been implemented in tiny IoT devices (FastGRNN and FastRNN). It is possible to fit FastGRNN in 1-6 kilobytes which makes this algorithm suitable for IoT devices, such as Arduino Uno.

Some of the well-known techniques considered for model size reduction include:

- *knowledge distillation*, whereby a small (easy to implement) model (student) is trained to behave like a larger trained neural network (teacher) while trying to preserve the accuracy of the teacher model, thus enabling the deployment of such models on small devices,
- steps such as quantization, dimensionality reduction, pruning, components sharing, etc. These methods exploit the inherent sparsity structure of gradients and weights to reduce the memory and channel occupation as much as possible,
- conditional computation reduces the amount of calculation by selectively turning off some unimportant calculations (for example with components shutoff, input filtering, early exit, results caching, etc.).

3.4 Distributed learning at the Edge

The computation resources of nodes at the Edge are limited in comparison to the cloud. Hence, the training of ML algorithms cannot be accomplished with a single edge node in many cases. There are several approaches to solve this problem by distributing the learning process.

The approaches can be divided into three categories. The first is data parallelism (*Figure 6.1*), which is about splitting the trainings data into smaller parts, training a model on each part and then implementing a model that combines the result of the other models. Due to the reduction of the amount of training data, the models are not as big and complex as a model trained on the whole dataset. In some cases, this means that one edge node can perform the entire training. In contrast to data parallelism, the second category, model parallelism (*Figure 6.2*), is about splitting the model into sub-modules and letting multiple nodes train each one of these modules on the same data. After the training is completed, the modules are combined again into one model. Offloading the training is the last category. Since the training and inference step have different requirements for computation power, it is possible to let a more powerful node take care of the training of the model and then to deploy the model for inference on a node with less resources for inference (*Figure 6.3*).

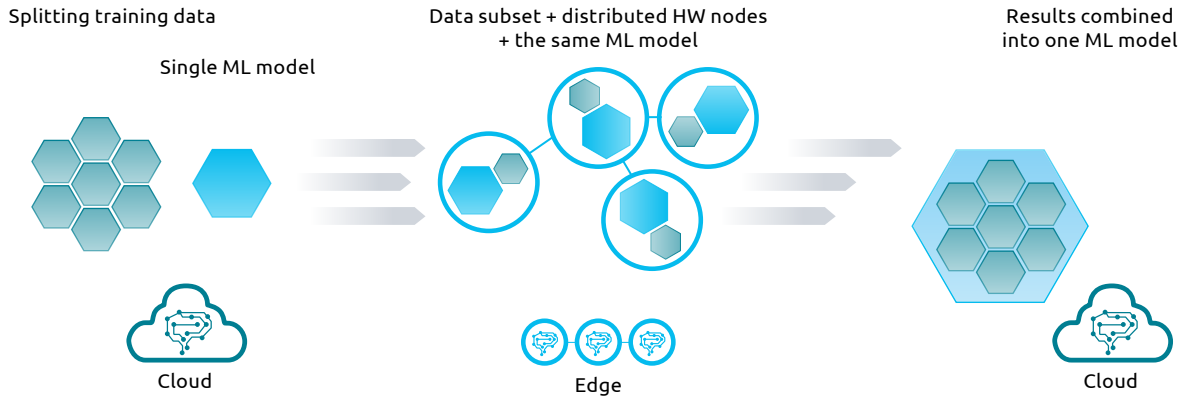


Figure 6.1: *Distributed Learning – data parallelism*

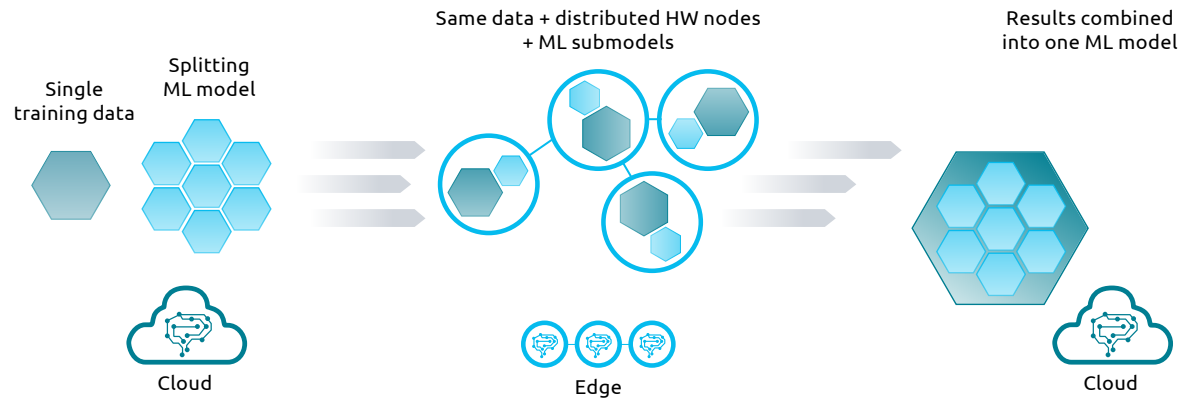


Figure 6.2: *Distributed learning: model parallelism*

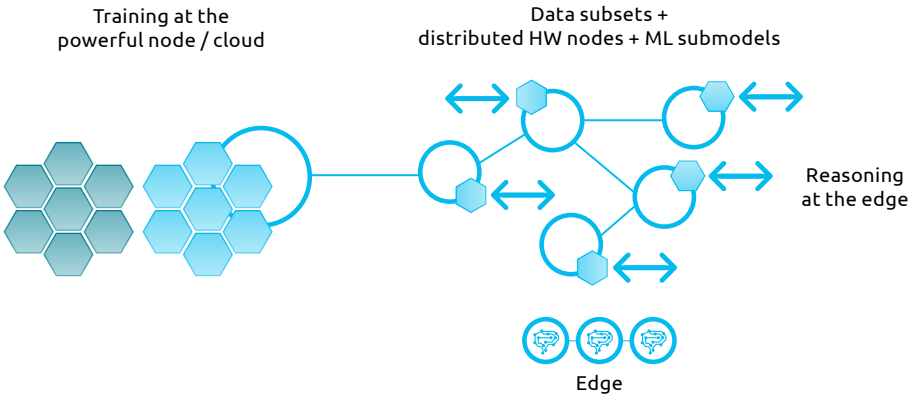


Figure 6.3: *AI model distribution and reasoning at the Edge*

A combination of data and model parallelism is *Federated learning*[33], which is based on training a series of local models on different devices, which are then combined in a central node for a global model upgrade. The central node is also responsible for the coordination between edge nodes. However, this approach involves trade-offs between model performance and communication overheads. The data in federated learning is split into smaller parts, which would be distributed amongst the nodes of the edge network, as shown in *Figure 6.4* below. Each node trains a separated model based on the data received, thus training a single part of the final DNN (Distributed Neural Network).

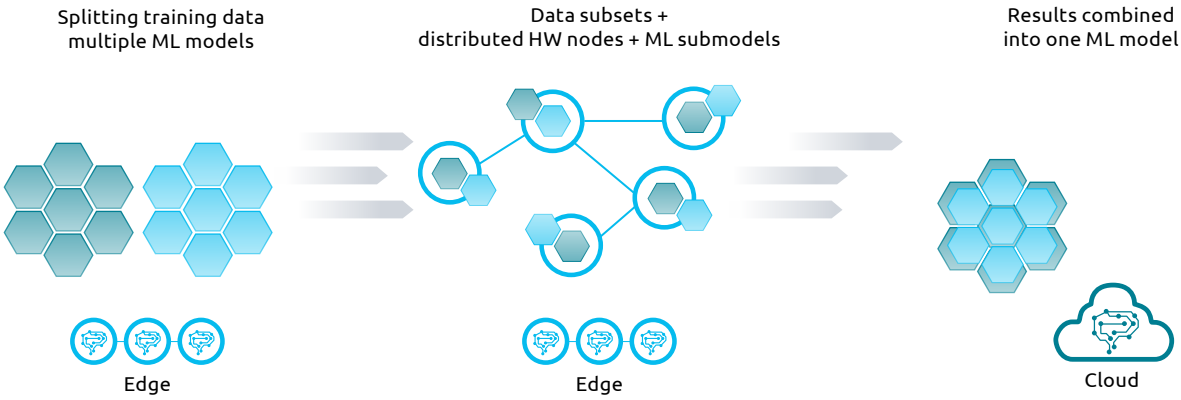


Figure 6.4: *Outline of the Federated Learning Approach*

3.5 Frameworks and platforms for AI at the Edge

Edge computing aims to bring high-performance computing capabilities and next-generation analytics powered by AI/ML, into hardware systems deployable at the edge. This requires the implementation of comprehensive intelligent edge frameworks and platforms whose development features the specific requirements and challenges spanning hardware, power efficiency, software, connectivity, flexibility and interoperability, and security.

Remaining major challenges include the end-to-end integration of connected systems, cloud endpoints, and third-party platforms or services, with a flexible embedded framework required to ensure maximum

use of data generated/collected over the long term. This is a necessary precursor to developing edge servers and processing solutions. Without a flexible edge-to-cloud integration platform and *supporting software/middleware*[34] libraries for *IoT edge gateways*[35] and other connected systems, solutions for high-performance edge computing cannot scale or adapt to the dynamic requirements of solution providers and end users. Other major challenges are represented by cross-platform flexibility (e.g. usable on both Android OS and Linux OS), dynamic parallelisation of the computational tasks, model compression, and end-user customization. To address the challenges for data analysis of edge intelligence, computing power limitation, data sharing and collaborating, and the mismatch between the edge platform and AI algorithms, Zhang et al. introduced an Open Framework for Edge Intelligence (OpenEI), which is a lightweight software platform to equip the edge with intelligent processing and data sharing capability[36].

The goal of OpenEI is that any hardware, ranging from Raspberry Pi to a powerful Cluster, will become an intelligent edge. Meanwhile, accuracy, latency, energy, and memory footprint, will have an order of magnitude improvement compared to current AI algorithms running on the deep learning package.

The framework includes:

- a Package Manager, which works as a running environment for AI algorithms on the edge platform, supporting both inference tasks and model re-training,
- a Model Selector, which is designed to find the most suitable models for the specific edge platform based on users' requirements in terms of accuracy, latency, energy, memory, etc.
- RESTful API, which is used for communication with cloud, other edge devices, and IoT devices

3.6 Orchestration of AI between cloud and edge resources

The implementation of complex AI-based applications like self-driving cars is very difficult due to limited resources of edge devices. Hence, these kinds of systems are often distributed between the edge and the cloud. It means often that the time-sensitive part of processing is implemented at the Edge and parts that can take more time are executed in the cloud. A specific example for such a procedure is the "Big-Little approach" by E. De Conick et al.[37]. They proposed to split a classification problem into a smaller part with a limited number of high priority classes and a larger part including all other classes. Afterwards, a model is trained for each part. Due to the reduction of number of classes, the size of the model handling the smaller part of the problem is reduced allowing it to be deployed on lower powered edge devices. In contrast, the larger model is deployed in the cloud or on an edge device with high amount of computation power as depicted in *Figure 7*.

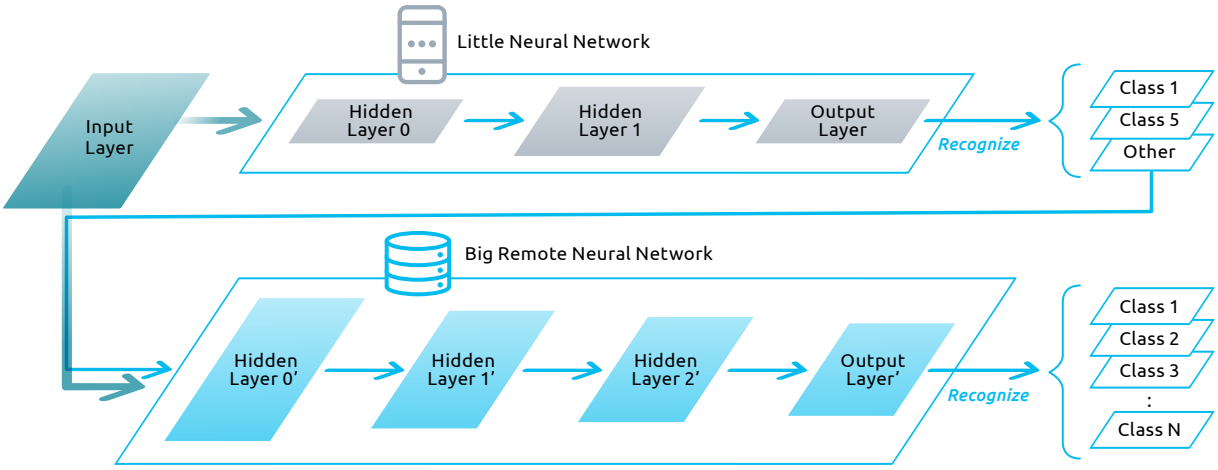


Figure 7: *Architecture of Big-Little neural network: the little neural network only classifies a subset of the output classes, and can be executed locally with limited CPU power. When the little neural network cannot classify the input sample, a big neural network running in the Cloud can be queried.*

Another distribution approach is to focus on the distribution of the processing of data. This means that as the data travels from its source at the Edge to the cloud each of the intermediate nodes that it passes perform a small share of the processing task until the final result of the processing pipeline is obtained in a cloud server. An implementation of this approach was proposed by S. Teerapittayanon et al.[38]. They exploited the fact that only the first few layers of a DNN are required for general processing to distribute one DNN over edge and cloud nodes. Furthermore, this implementation also introduces exits points that allow the termination of processing when the results are sufficiently accurate, are required earlier due to time constraints, or cannot be processed further due to a node failure. The deplyoment of this approach is presented in *Figure 8*.

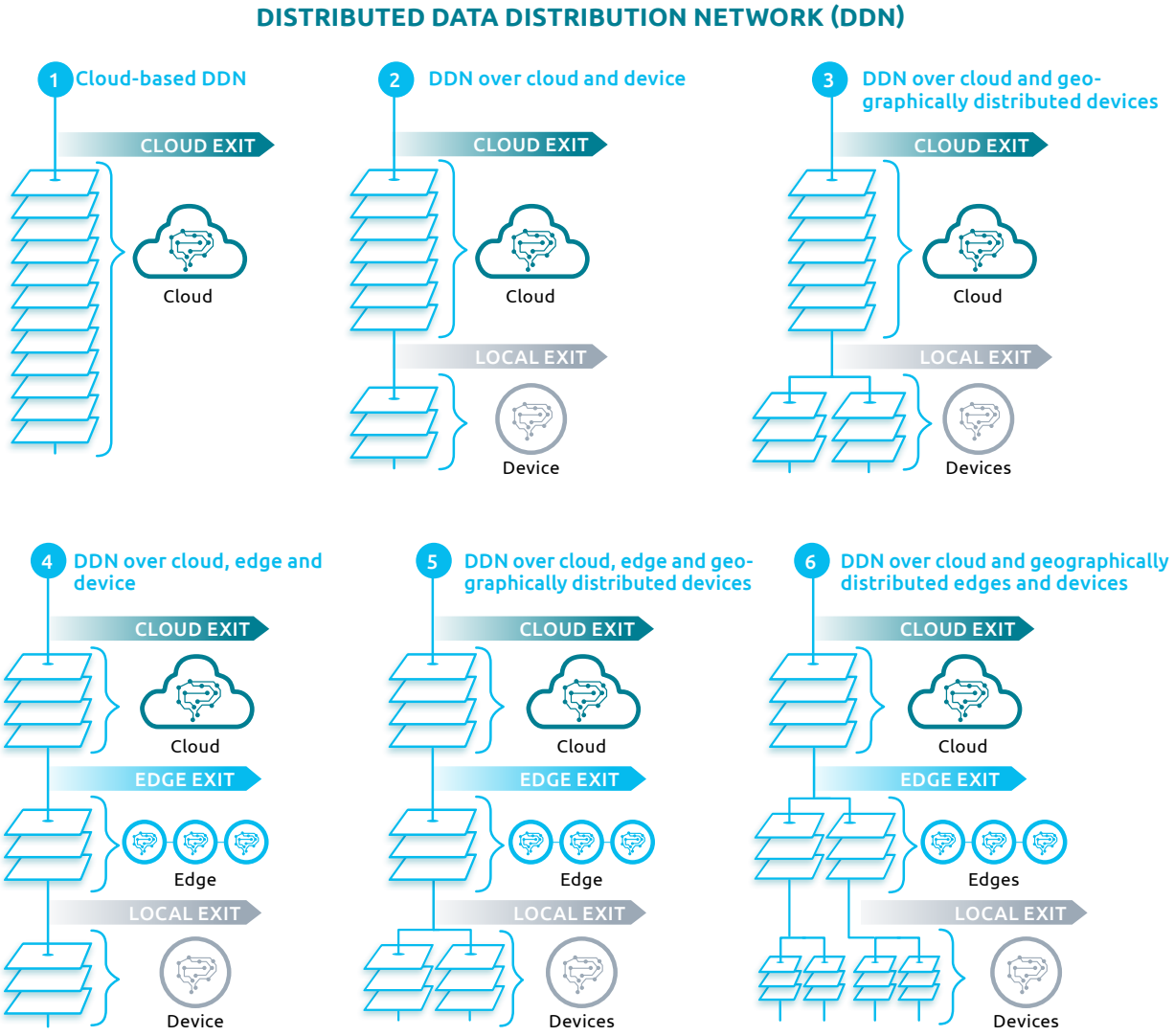


Figure 8: *Overview of the DDNN architecture. The vertical lines represent the DNN pipeline, which connects the horizontal bars (NN layers): (1) is the standard DNN (processed entirely in the cloud), (2) introduces end devices and a local exit point that may classify samples before the cloud, (3) extends (2) by adding multiple end devices which are aggregated together for classification, (4) and (5) extend (2) and (3) by adding edge layers between the cloud and end devices, and (6) shows how the edge can also be distributed like the end devices.*

3.7 Hardware-software co-design for AI at the Edge

One of the most important challenges in the implementation of AI at the Edge is to be able to offer scalable solutions and yet meet diverse application needs, in terms of:

- end-users' varying context (e.g. job to be done),
- individual end-users' characteristics (such as demographics),
- latency expectations,
- available battery power and computational power within the device.

To address these challenges, it is quite important to ensure that both the hardware and software adapt to the dynamic context at the Edge and devices' state. Here, *hardware-software co-design* helps to ensure that this adaptation and personalization happen seamlessly.

While, in general, AI models have been designed with a top-down design flow, mainly focused on achieving the highest possible accuracy and performance, assuming that the HW will deliver the computational tasks required. This approach ignores the limitations present in the deployment of intelligent systems at the Edge. Instead, AI models should be built bottom-up with adequate understanding of the hardware constraints. In order to provide an optimized solution it is most important that AI models and the associated HW are developed simultaneously.

A good example of this co-design approach is that some smart sensors include self-learning AI together with other non-AI signal processing functions. As the sensor's co-processor is capable of executing context-sensitive firmware on-demand, the device can switch between AI and non-AI firmware depending on the need. This solution can thereby reduce electronic-waste by having specialized hardware for AI and minimise overall bill of material cost. Additionally, the co-design of software and hardware helps to extend, or easily integrate, further physical and virtual sensors (e.g. magnetometer, pressure sensors, inertial sensor, etc.) as additional external inputs. This enables faster and more robust learning from an expandable list of input sources, chosen according to edge application, as opposed to pre-programmed (AI) solutions with a fixed number of physical inputs and without a built-in learning function. As the self-learning AI function executes on the sensor's co-processor, the overall system power and memory requirements are extremely low in comparison to other non-edge AI systems.

In summary, as highlighted in the previous paragraphs, H/W aspects, models, and communication platforms are inter-linked when developing a system working at the Edge. Hardware-software co-design of edge-AI systems provides a path to the execution of a wide variety of applications (AI and non-AI included), whilst having the capability to adapt to the application needs on-demand.

4 Future Challenges and Trends

The development and deployment of a secure and trustworthy Edge AI will require a wide number of challenges to be addressed and solved.

4.1 Trust and explainability

AI algorithms, and especially deep neural networks, are often considered as black boxes. The decision-making process of typical ML algorithms is not always transparent, and usual data models based on NN do not represent the characteristics of the process to be represented. Furthermore, their internal computations present a black-box and are not easily understandable for humans. The drawbacks of such algorithms include:

- any bias within the training data is potentially transferred to the algorithm and remains undetected,
- users may not trust their predictions,
- and that they lack robustness in operational environments.

Explainable AI aims to provide insights into the internal decision-making process of machine learning algorithms. Using these insights, algorithms can be developed whose predictions are not only correct but right for the right reasons[39].

4.2 Re-learning

Acceptance and market uptake of products and services are directly dependant on trust into offered solutions. That is particularly manifested in emerging automotive applications, such as Driving Automation, which are heavily reliant on edge AI. Hence, it is of utmost importance that a common approach to AI is based on trust and excellence.

Considering the importance of edge AI, there is a need for commitment to consider the impact of edge AI throughout its lifecycle. To that extent, the developed algorithms must be kept up to date and performant on new data, with the ability to integrate external sources through re-training. In addition to meet the requirements and defined metrics that indicate the training state of the AI system, the re-training must also consider any consequence it may have on other components or the system itself. The implication is that rather than having to spend time and resources on re-training from scratch, to incorporate slightly different insights, the re-training should focus on creating more generic models. The aim is to permit improvements in performance through a quick re-training of an edge AI model that has already been trained using previous data sets. Equally, the re-training of one model should not compromise the performance of other components within the system (or other systems within a system of systems). In simple terms, the re-training must enable improved performance through exploitation of new data and in parallel it must not negatively impact its surroundings. Additionally, changes in calibration (e.g. of sensors or actuators) should be permitted without the need to retrain the edge AI.

4.3 Security and adversarial attacks

In distributed learning, a communication overhead is introduced in order for the edge platforms and the system aggregator to transfer data during training and inference. When compared to data processing in large central data centres, data produced on resource-constrained end devices in a decentralized distributed setting is particularly vulnerable to security threats and the necessary level of protection against such risks should be considered carefully for specific applications.

Figure 9: *A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we physically place a sticker targeted to the class "toaster" on the table (bottom photograph), the photograph is classified as a toaster with 99% confidence (bottom plot). See the following video for a full demonstration: https://youtu.be/i1sp4X57TL4 (Source: https://arxiv.org/pdf/1712.09665.pdf, Foto: Pixabay)*

An example of a security threat debated in the recent years is adversarial attacks. Adversarial attacks describe the use of erroneous data to manipulate the results of AI algorithms, especially of neural networks. In the context of image or video classification, attacks are done by designing specific noises, colours, lighting, or orientation patterns, which are then integrated in the corresponding data. An example of the so-called "adversarial patches attack" was presented at NIPS 2017 – Conference on Neural Information Processing Systems. After their generation these patches can be placed anywhere within the field of view of the classifier and cause the classifier to output a targeted class. In *Figure* 9 above, a banana is correctly classified as a banana. Placing a sticker with a toaster printed on it is not enough to fool the network and it continues to classify it as a banana. However, with a carefully constructed "adversarial patch", it is easy to trick the network into thinking that it is a toaster. This patch attack is especially difficult as these patches can easily be distributed after their creation.

Adversarial attacks exist for other kinds of AI-based data processing, e.g. audio or LiDAR (Light Detection And Ranging). However, to date, these areas are not as well investigated as attacks on image or video classifiers.

Further research is required to increase the security, privacy, and robustness of edge AI by reducing the overhead, or by adopting novel approaches such as clustered federated learning or federated distillations.

4.4 Learning at the Edge

Training artificial neural networks at the Edge remains a challenge. Work has been done to optimize inference at the Edge by optimizing algorithms and accelerators for low precision, low memory footprint and feed-forward computations. However, an additional re-training phase of an artificial neural network can undo part of those optimizations as higher precision is needed to enable the iterative approach typically used and more storage is needed to keep track of the intermediate data required. Also, the frequent weight updates during training can pose additional challenges regarding energy efficiency as well as reliability.

As such, neuromorphic-based architectures hold potential, as they allow on-line learning to be built in by modelling plasticity. Plenty of challenges remain to achieve this goal as it is difficult to make a single synapse and neuron device that allows the capture of a very wide range of time constants.

Another approach for edge learning is to implement a pre-trained neural network for inference but permit adaptation of the final network layers to tune classification or detection, this approach is called transfer learning.

4.5 Integrating AI into the smallest devices

Recently a number of tools have been developed with the goal of implementing AI models which could fit the memory available in edge platforms.

As an example, tinyML is about processing sensor data at extremely low power and, in many cases, at the outermost edge of the network. Therefore, tinyML applications could be deployed on the microcontroller in a sensor node to reduce the amount of data that the node forwards to the rest of the system. These integrated "tiny" machine learning applications require "full-stack" solutions (hardware, system, software, and applications) plus the machine learning architectures, techniques, and tools performing on-device analytics. Furthermore, a variety of sensing modalities (vision, audio, motion, environmental, human health monitoring, etc.) are used with extreme energy efficiency (typically in the single milliwatt, or lower, power range) to enable machine intelligence at the boundary of the physical and digital worlds.

Tensorflow Lite (TFLite) was created specifically for this purpose, it proposes a set of tools that help programmers to run AI models on embedded, mobile, and IoT devices. A typical workflow will involve the definition of the AI model in Keras/Tensorflow, followed by the conversion of the model from Keras to TFLite, and the final compression of the model (for example, via post-training quantization) to further decrease the overall footprint. Many tinyML implementations actually use TFLite under the hood.

With the increase in dedicated hardware for machine learning, an important direction for future work is the development of compilers, such as Glow, and other tools that optimize neural network graphs for heterogeneous hardware or train and handle specialized technologies and algorithms.

4.6 Data as a basis for AI

Data is the fundamental piece behind ML/AI. However, one of the major problems when developing AI solutions can be the lack of sufficient data to achieve the required performance in a specific application. In recent years several techniques have been considered to deal with this problem in the context of cloud-based solutions; for example, by using semi-supervised learning (to take advantage of the large amounts of unlabelled data generated by edge devices), by using data augmentation (via Generative Adversarial Networks (GANs) or transformations), or by transfer learning. These have become cutting-edge methods deployed to improve the overall performance in AI models. However, the adoption of these techniques in edge computing still needs to be thoroughly investigated.

Moreover, edge systems need to interact with various types of IoT sensors, which produce a diversity of data such as image, text, sound, and motion. Edge analytics should be able to deal with those heterogeneous environments and adapt to be multimodal allowing learning from features collected over multiple modalities.

4.7 Neuromorphic technologies

Neuromorphic engineering is a ground-breaking approach to the design of computing technology that draws inspiration from powerful and efficient biological neural processing systems. Neuromorphic devices are able to carry out sensing, processing, and control strategies with ultra-low power performance. Today, the neuromorphic community in Europe is leading the State-of-the-Art in this domain. The community includes an increasing number of labs that work on the theory, modelling, and implementation of neuromorphic computing systems using conventional VLSI technologies, emerging memristive devices, photonics, spin-based, and other nano-technological solutions. Extensive work is needed in terms of neuromorphic algorithms, emerging technologies, hardware design and neuromorphic applications to enable the uptake of this technology, and to match the needs of real-world applications that solve real-world tasks in industry, health-care, assistive systems, and consumer devices. It is important to note that "neuromorphic" is most commonly defined as the group of brain-inspired hardware and algorithms.

Parallel to the advancement in neuromorphic computing, the underlying computation of such technology gets increasingly complex and requires more and more parameters. This triggers further development of efficient neuromorphic hardware designs, e.g. the development of neuromorphic hardware that can tackle the well-known memory wall issues and limited power budget in order to make such technology applicable on edge devices. The emerging memory technologies provide additional benefits for neuromorphic solutions, especially memory technology that can allow us to perform computation directly in the memory cells themselves instead of having to load and store the parameters, inputs, and outputs into computation cores.

Such technology, coupled with the properties of neuromorphic computing, delivers many benefits. Firstly, DL and spiking neural networks (SNN) parameters are often fixed and/or modified very seldom. This matches the capability of emerging non-volatile memories where write accesses are typically one or two orders slower than read accesses as the number of memory writes required is lower. Secondly, most computations are matrix addition and multiplication. This operation can be mapped efficiently in memory arrays. Thirdly, inference of such neuromorphic networks can be optimized for low-bit precision and coarse quantization without sacrificing the quality of the network outputs. Some tasks, such as classification, are proven to be good enough even when networks are optimized to binary and/or ternary representation. This provides an excellent opportunity as the underlying operation can be simply replaced by AND/XOR logic. Fourthly, neural networks are robust to error. Thus, process variations on the emerging memory technologies do not limit their capability to compute and/or and load/store in the networks. These benefits can be achieved by in-memory compute technology using emerging memory technologies.

4.8 Meta-learning

In most of today's industrial applications of deep learning, models and related learning algorithms are tailor-made for very specific tasks[40][41]. This procedure can lead to accurate solutions of complex and multidimensional problems but it also has visible weaknesses[42][43]. Normally, these models require an enormous amount of data to be able to learn how to correctly solve problems. Labelled data can be costly as it may require the intervention of experts or not be available in real-time applications due to the lack of generation events.

A question can therefore arise: in addition to having the correct formulation and the descriptive data for the problem, is it possible not only to try to solve it but also to learn how to solve it in the best way? Therefore: "is it possible to learn how to learn?" Precisely on this question, the branch of machine learning, called **meta-learning** (Meta-L), is based[45][46].

In Meta-L the optimization is performed on multiple learning examples that consider different learning objectives in a series of training steps. In **base learning**, an inner learning algorithm, given a dataset and a target, solves a specific task such as image recognition. During **meta learning**, an outer algorithm updates the internal algorithm so that the model learned during base learning also optimizes an outer objective, which tries, for example, to increase the inner algorithm's robustness or its generalization performance[47].

This two-step iterative approach is resulting in successful solutions to problems where few labels or, in general, little data is available, as the highest level of information is extracted thanks to the formulation of the optimization problem itself. Intelligent extraction of information, by addressing the problem from a general point of view can also lead to the ability of the inner algorithm to handle new situations quickly and with little data available with a robust approach[48].

Exactly for the reasons listed above, Meta-L is gaining significant attention in Edge AI, where the new data collected can be immediately processed and fed to the algorithms to increase the robustness of the model and generalisation of new tasks that may be useful for systems, even in the deployment phase. Looking at the advantages of Meta-Learning and the possibility of using it together with Edge computing to increase its benefits, provides a good outline of how this branch of ML can soon find concrete uses in the most varied application scenarios[49].

4.9 Hybrid modelling

Data-based and knowledge-based modelling can be combined into hybrid modelling approaches. Some solutions can take advantage of a-priori knowledge in the form of physical equations describing known causal relationships in the behaviour of the systems or by using well known simulation techniques. Whereas dependencies not known a priori can be represented by many kinds of machine learning methods using big data based on observing the behaviour of the systems. The former type of situation can be seen as *white box* modelling as the internal states possess a physical meaning, while the latter is referred to as *black box* modelling, using just the input-output-behaviour, but not maintaining information on the internal physical states of the system. However, in many cases, a model is not purely physics-based nor purely data-driven, giving rise to grey box modelling methods that can be formulated[50]. The assignment of models to the scale varies within the literature: For instance, a transfer function can be derived from physical considerations (white), identified from measurement data with a well-educated guess of the model order (grey) or without (black).

Approaches for combining machine learning and simulation, by simulation-assisted machine learning or by machine-learning-assisted simulation and combinations are described by von Rueden et al. in "Combining Machine Learning and Simulation to a Hybrid Modelling approach: Current and Future Directions"[51] and in "Informed machine learning – towards a taxonomy of explicit integration of knowledge into machine learning."[52] advantage of hybrid modelling is avoiding the necessity of learning a-priori the behaviour of systems from huge amounts of data, if they can be described by simulation techniques. Also, in the case of missing data, hybrid modelling is a possible approach[53].

A practical example of combining physical white-box modelling and machine learning to improve a model for the highly non-linear dynamic behaviour of a ship, described by a set of analytical equations has been recently investigated by Mei et al.[54]. Another example is hybrid modelling in process industries[55].

4.10 Energy efficiency

Reducing energy consumption is a general goal, not only, but especially for smart systems providers to address the challenges of global warming and enable a higher degree of miniaturization of intelligent devices. For a long time power reduction has been a challenge in micro and nano electronics and also a target for all AI applications, regardless of whether data is processed in the cloud or at the edge. But at the edge, this target is especially important as applications usually have only limited power resources available. They often have to be battery powered or even use energy harvesting.

Special energy-efficient neural network architectures have been investigated[56]. Not only is the hardware crucial for low-power AI applications, but also the implemented methods and models have great influence on the energy consumption. This has been examined for the example of computer vision[57].

Moving away from traditional von Neumann processing solutions and using dedicated hardware[58] allows for additional power reduction. Even more can be achieved with neuromorphic architectures[59].

The “ultimate benchmark” in power consumption for artificial intelligence would be the “natural intelligence” in form of the human brain, which has 86 bn. neurons[60] and approximately 10^{14}–10^{15} synapses[61] with an energy consumption of less than 20W, based on glucose available to the brain, or only 0.2W, when counting the ATP usage instead of glucose[62]. Current GPU based solutions with that complexity are far from this energy efficiency. There is obviously plenty of headroom for further development.

5 Milestones for AI at the Edge in Smart Systems

Edge AI is a key technological area that is ending the pure dominance of the cloud in the data analytics world. As shown by the numerous scenarios and contexts reported in this white paper, Edge AI technology is poised to disrupt a wide variety of industries because of the huge advantages introduced, such as increase real-time performance, improved energy efficiency, improved security and privacy etc.

The evolution of a new generation of edge intelligence systems will take place during the next 5–15 years, with the completion of different technological steps supporting the development of new devices, technology and applications.

However, there are still several challenges that have to be addressed:

- the development of new algorithms and applications,
- the development of neuromorphic-based chips and new specialized computing platforms and their integration with classical systems,
- the development of efficient and automated transfer learning to support federated learning as well as the optimization of neural networks from general-purpose to application-specific scenarios,
- the implementation of new tools and frameworks allowing (semi)-automatic design exploration as well as an automatic generation of deep networks architecture,
- the development of open architecture (based on opens source SW, open data, open edge platforms, open HW) allowing fast turn-around deployment,
- the implementation of energy and cost-efficient AI training on the Edge and security, privacy, and explainability.

We expect a significant growth in these research fields that will address the main challenges identified in this white paper.

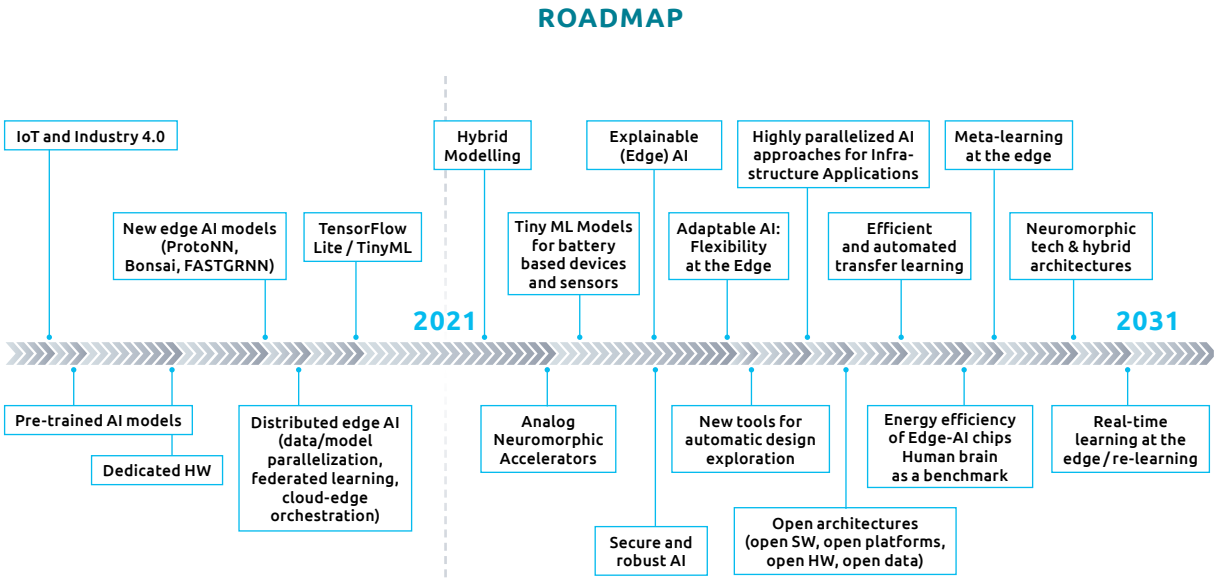


Figure 10: *Roadmap for AI at the Edge in Smart Systems*

6 Policy Recommendations

6.1 Sustainable business model innovation

While the edge AI building blocks are contributing towards expanded functionality and more dependable Cyber Physical Systems (CPS), it is the exploitation of these results that is going to create impact. That impact can be maximised if solutions to the technical challenges are capable of improving end-user acceptance and consequently lend themselves to increased usage. It is the documented use cases (such as those of *chapter 2)* that have a potential of determining the suitability of the technological advances for entry into the mass market. Such commercialisation is driving the formation of appropriate go-to-market strategies. The selected strategy must take into consideration a wide range of stakeholders who are crucial for the acceptance of new solutions. The resulting cooperation is also driving cross-fertilisation across the industrial domains. The anticipated impact is based on the ability to improve real-time decision making and transforming the way that business is done, hence offering opportunities for sustainable implementation through expanded capabilities, gained efficiency and improved business processes. These opportunities must be taken immediately as the fast-paced technological evolution leaves little room for hesitation when integrating AI into existing businesses. The integration should consider the following principles:

- **Objective:** The starting point for consideration should be the required outcome for existing or new business models, rather than the technology and its integration.
- **Start Small:** As the off-the-shelf solutions are becoming more common and easier to leverage, it is crucial to perform incremental (learning infused) steps. Small, but scalable, applications aid understanding of the user needs for a specific domain. This is then followed by scaling up to further domains and solutions.
- **Continual Upgrades:** As algorithms are adapting their performance based on the training data, it is less likely for it to be applied to a full system, as customisation would be complex and the potential benefit may not be worth the effort. Hence, performing small steps and building on the generated solutions is a reasonable way forward. That may, but does not have to, rely on the principle of Minimum Viable Product (MVP).
- **Open Collaboration:** The small start and additional steps must also consider the need for the very wide range of skills required and the fact that there are few stakeholders who are capable of implementing full solutions. Hence, it is highly advisable to employ open innovation and collaborative projects where partners with complementary core competencies can join forces to create solutions for the common good while supporting the interests of individual organisations.
- **Evolving Innovation:** the progress could benefit from the evolution of innovation. By building benefit-yielding functionalities one must monitor improvements to the existing business activity. It is normal to encounter some resistance to change due to the apparent lack of immediate returns on investment. So, it is crucial to highlight incremental improvements and the long-term vision.
- **Job Market Transformation:** one should consider that this transition period is removing the need for certain job roles, (e.g. maintenance) and at the same time creating new employment opportunities[63]. In such an environment there is an implicit need for (re)training to take advantage of the transforming market and to keep up with the changes in the job market.
- **Refocus:** There is a paradigm shift from the focus on the traditional solutions (such as solely production-based) towards the provision of services. This is especially evident in the automotive sector.
- **Data:** Considering that data fuels edge AI's superior performance, one must define what value AI needs to provide. That answer should be followed with the definition of the required, available and missing data.

- **Customisation/personalisation**: Improved customisation is resulting from the edge-AI learnings from the available data. In terms of business opportunities, this implies leveraging data can improve the understanding of customer behaviour and their needs in order to evolve a deeper and more personal approach to customer engagement. Such customisation generally improves user experience and hence lowers resistance to the new technologies.

An emerging issue from the above set of principles, which is especially related to the transformation in the job market, is that of emerging gaps in the ecosystem, as depicted in *Figure 11*. In principle, many of the edge AI-related assets already exist. However, the interconnections are often missing because of development being undertaken in silos. These gaps must be bridged and, in turn, enhance the potential for business model innovation. If that innovation process is continuously driven it increases the probability of sustainable businesses. The sustainability is further underpinned by a mixture of different skill sets in combination with the industrial domain knowledge i.e. the technology providers and industrial users are mutually benefiting from the cooperation. The inevitable benefits are the enabled learning for all, through cross-fertilisation and creation of a competitive advantage over generic solutions.

By bridging the gaps within the ecosystem, it is inevitable for the new opportunities to be created and complement the existing sources of revenue (*Figure 11*). Upon identifying the new sources, organisation should focus on achieving the new revenue mixture (e.g. use cases, technologies, products, services, customers/users, partners etc.) to deliver the envisaged growth opportunities.

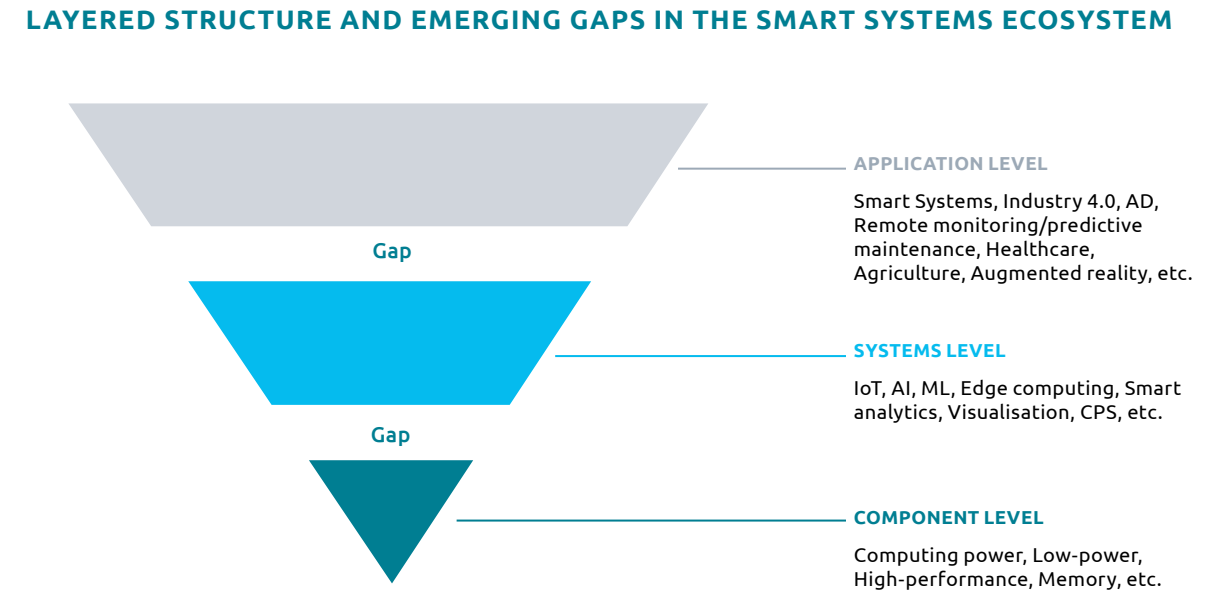


Figure 11: *Layered structure and emerging gaps in the ecosystem*

6.2 Our vision – cross domain technology stack

The technology stack for data-driven applications[11] is continually evolving in relation to the European communities. *Figure 12* depicts grouped components of the stack according to functionality i.e. centralized computing, connectivity and edge applications. The tight collaboration with further stakeholders of the "digital stack" is required to finally provide a consistent technology stack. As highlighted in *section 1.3*, value creation results from deployment of tailored end-to-end solutions for specific application domains. The deployment across different

domains helps identify the common requirements (and resulting technologies). It also enhances understanding in terms of how to tailor this common basis to domain specific activities.

As a result of this, the "competence-focused" way of thinking must be re-organised into an "end-to-end solution" way of thinking. New ways to interact with the relevant (European) communities are required to integrate the relevant stakeholders effectively and efficiently in this process and, finally, contribute to the European industrial digital transformation.

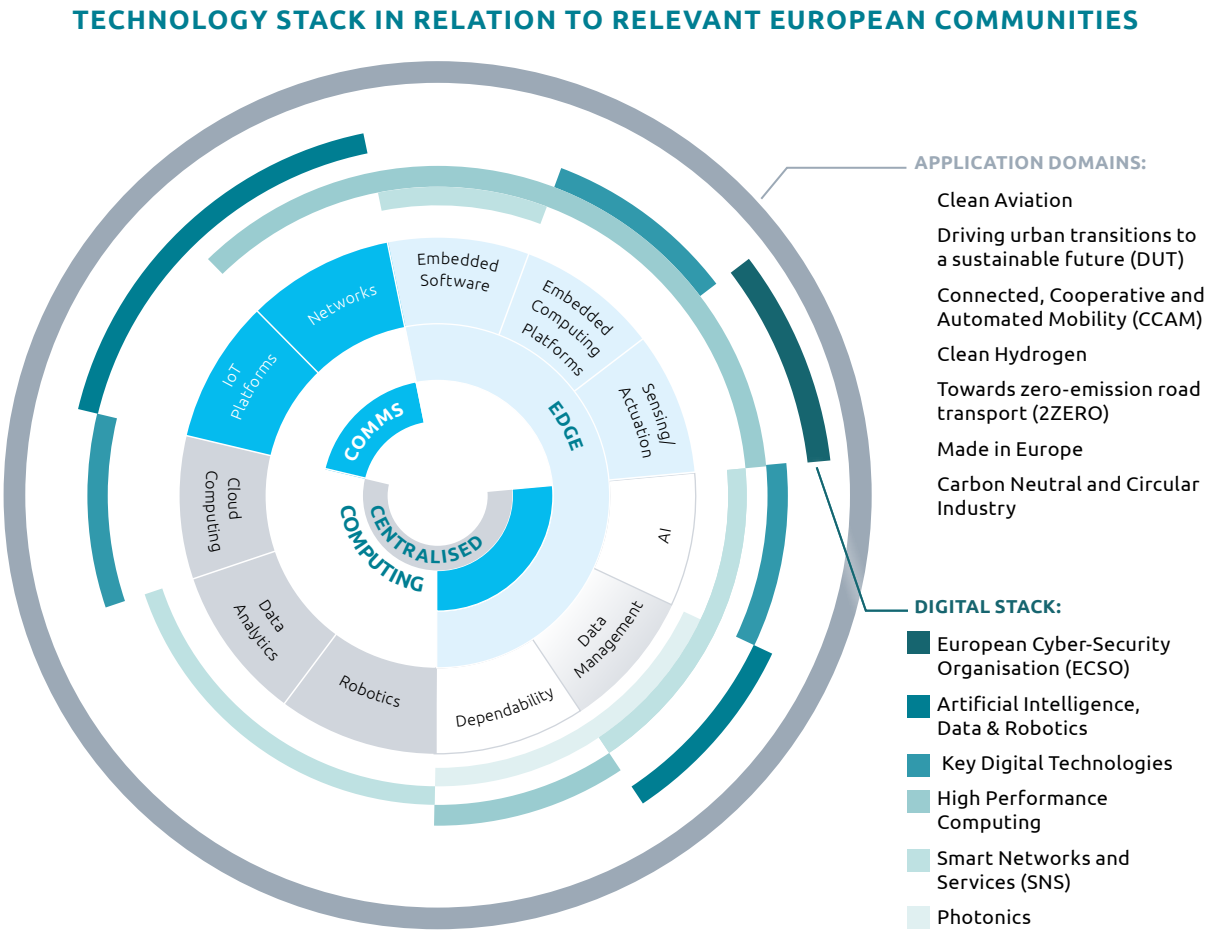


Figure 12: *Technology stack in relation to relevant European communities*

6.3 Common standards

One of the major problems faced by data scientists when using data from external sources is to understand how the data was collected, and what it represents. Data management standards, such as CRISP-DM, need to be enforced on a wider scale with the goal of reducing production time and making the development of AI solutions simpler. Moreover, data sharing on platforms should be promoted as much as possible to ensure that academia and industry can leverage the full potential of the collected data, as also currently proposed by several global funding bodies.

Several efforts have recently been deployed with the goal of defining a common edge computing view worldwide. For example, ISO/IEC TR 23188 aims to describe edge computing and the significant elements which contribute to the successful implementation of edge computing systems. It is based on an emphasis on the use of cloud computing and cloud computing technologies in the context of edge computing, including the virtualization of compute, storage and networking resources. Moreover, ISO/IEC TR 30164 takes a view of edge computing from the point of view of IoT systems and the IoT devices which interact with the physical world.

For example, the upcoming MPEG-7 Part 17 standard (ISO/IEC 15938-17 NNR) will define tools for compression of neural networks for multimedia applications and representing the resulting bitstreams for efficient transport. NNR targets a compression efficiency over 95% without degrading classification quality.

Finally, eighteen organizations (including Huawei, Analog Devices, Arm, Bombardier, Fraunhofer Institute for Open Communication Systems – FOKUS, German Edge Cloud – GEC, German Research Center for Artificial Intelligence – DFKI, IBM, Intel, National Instruments, Renesas Electronics, Schneider Electric, etc.) have signed a cooperation agreement to form the European Edge Computing Consortium (ECCE). ECCE aims to provide a comprehensive edge computing industry cooperation platform with the aim to create a standard reference architecture and technology stack that can be deployed across smart manufacturing, other industrial IoT applications, and network operators. The goals of the initiative include the specification of a reference architecture for edge computing (ECCE RAMEC), the development of reference technology stacks (ECCE edge nodes), the identification of gaps and the recommendation of best practice. These are based on evaluating approaches within multiple scenarios (ECCE Pathfinders) and the synchronization with related initiatives, standardisation organisations and the promotion of the results.

6.4 Heterogeneous approaches, multiple vendors

Many different AI development frameworks are in use today, often originating from one of the major cloud providers. These frameworks are vertically integrated and seldom show portability towards edge devices. The best-supported exchange format is ONNX (Open Neural Network Exchange format). This is however still a high-level description of the neural network and requires a complete tool flow to allow the execution on an embedded neural accelerator. Hardware vendors need to support several software flows, as there is no de facto standard. There is a strong need for standardised interfaces / API definitions to describe the hardware capabilities etc.

Having formalised interfaces between the different implementation steps will enable hardware and software vendors to provide parts of the tool flow (compiler, deployment tools, hardware-aware code transformations) without the overhead of developing multiple interfaces. This ensures that AI developers can target multiple platforms with relative ease. A good example is the Apache TVM flow, which allows mapping from ONNX (Open Neural Network Exchange) to compiled code, and provides interfaces and intermediate representations with which third-party tools can interface. Promoting projects focused on inter-operability and standard tool flows and interfaces has the potential of opening the market for both large and small companies

6.5 Education and network building

Despite the importance that edge AI is showing at a European level in a number of industries, companies are facing a shortage of highly-skilled scientists in this field. This is also due to the lack of specific training, modules or courses in third-level education which could encompass all the multidisciplinary aspects of edge AI (e.g. hardware platforms and electronics, firmware development, embedded systems, data mining and machine learning for cloud platforms and resource-constrained devices, communication protocols, etc.).

Moreover, it is evident that there is a lack of networks focussed on this topic, which could strengthen and connect the small and fragmented community of researchers and engineers in academia and industry.

6.6 Data collection, testing and experimentation facilities for AI at the Edge

Scientific experiments are generally considered to be controlled studies aimed at enhancing understanding of relationships between the cause and effect. The focus is on data collection through either surveys or, more objectively, instrument-based methods. When it comes to edge AI applications, the distinctions may become blurred. Considering the black-box property of AI solutions, this relationship between the cause and effect is upgraded to a higher level of abstraction, which poses a greater challenge in terms of analysis and understanding. Additionally, edge AI solutions frequently place humans in the control loops. This human-centric approach imposes the need for mapping and alignment of objectively acquired sensor data with the subjective estimates, as perceived by the participants. This is emphasised in automotive applications e.g. driving automation, where control functionality must take into account the passenger stress and comfort levels.

Some of the direct implications of the edge AI data properties on experimentation for emerging applications are:

- **Regulatory/certification:** While deemed as one of the most promising contributors for the development of highly automated driving, provision of holistic edge AI algorithms as a replacement of the human decision making is unlikely in the near future[64]. The main stumbling block is the non-compliance of AI solutions with imposed regulations for safety-critical applications (e.g. automotive or avionics domains). A potential solution to this issue is seen in a conceptual shift from pure experimentation in standard test facilities towards a mixture of those tests with continual and holistic experimentation during the full life cycle of the offered assets.
- **User acceptance and trust:** As specified in section 2.4, a common approach to AI must be based on trust and excellence, as the black box issue feeds into end-users' fears and lack of trust in decisions made by edge AI-based solutions. As trust is the key component for acceptance of driving automation[65], an added focus on the demonstrations is a method of improving user acceptance. Hence, testing facilities and experiments must be adapted to facilitate increased exposure of the novel technology.
- **Cyber-security:** As the edge AI utilises connectivity, the resulting highly connected networking exposes vulnerabilities and amplifies the attractiveness level for cyber-attacks. That is further contributed by the usage of certain sensors. The elimination of consequent privacy protection issues and compliance with the GDPR are reliant on identification and mitigation of risks posed by potential cyber-attacks. These challenges must be considered when testing and experimenting with the novel solutions, or else there is a significant potential for breach of privacy.

7 Summary

In this white paper experts from the smart systems integration community of EPoSS present their views on the state-of-the-art and future technology milestones in the Edge AI domain. They have put forward a range of policy recommendations aimed at accelerating the development and deployment of Edge AI solutions in the coming decade.

The experts from the Automotive, Energy, Industry, Health, Agriculture and Smart Cities domains provided use cases to illustrate the huge potential for edge AI applications and their future benefits for our society. Some areas, such as AI specific hardware/accelerators, distributed learning for edge methods and algorithms, platforms and frameworks for software and hardware co-development, have already reached a high maturity level for "real world" applications. However, there are still security, safety and trustworthiness challenges to be solved to enable the full potential of AI at the Edge. These challenges may apply to AI in general but many have additional specific needs for applications at the Edge. Future research and applications will be driven by achieving technology milestones such as: implementation on the smallest devices, high quality data, meta-learning, neuromorphic computing and other novel hardware-architectures.

Finally, the experts compiled a set of policy recommendations as a framework for R&D&I projects to address objectives such as sustainability, energy efficiency, safety and security. These should guide the future research and development efforts. Lower energy consumption of both hardware and software algorithms will underpin these ambitious goals.

Critical will be cross-domain and cross-technology working that will allow cooperation between various vendors combining the best hardware and software know-how and technologies.

Realization of the vision described in this paper requires a set of concrete actions and coordinated effort by companies, academia and public authorities. In the face of strong international competition in this developing field, a fast exploitation of the broad range of available state-of-the-art technologies is of highest importance for Europe. The leadership in cloud computing is lost for European players – but AI at the Edge is still an open opportunity and a must for European smart systems providers (especially sensor companies) to remain competitive and maintain their strong position in these markets.

To achieve the industrial goals, the experts propose the following actions:

- Provide internationally compatible funding for academic research in AI at the Edge and for companies to address their R&D&I needs
- Make security, privacy, energy consumption and sustainability key attributes of European AI at the Edge solutions
- Strengthen R&D&I projects by enabling cooperation along and across value chains for both hardware and software experts in the field of smart systems and the AI and IoT community
- Address the new engineering and software development skills needed, in both AI hardware and software through support of cross-domain software and hardware education and network building for both academia and industry
- Provide incentives to build the European talent pool to maximise the impact of European initiatives
- Support the development of European data spaces, in order to collect and share high quality, trustworthy data as outlined in the European Data Strategy[66]

- Develop standards for distributed data exchange and machine learning models to complete the tool-chains
- Establish experimentation and test facilities for distributed data collection and software and hardware co-design
- Increase usability, acceptance and safety based on considered regulation and efficient certification

Further steps require in-depth analyses of the European ecosystem and a closer collaboration across all actors in the field of IoT, AI and the smart systems and electronics community.

References

[1] https://iot-analytics.com/rise-of-iot-semiconductor/?utm_source=IoT+Analytics+Master+People+List&utm_campaign=4f683a3da9-The+rise+of+IoT+semiconductors+BLOG&utm_medium=email&utm_term=0_3069fbcae4-4f683a3da9-345823361

[2] https://stateoftheedge.com/reports/state-of-the-edge-report-2021

[3] https://iot-analytics.com/iot-edge-computing-what-it-is-and-how-it-is-becoming-more-intelligent

[4] https://stateoftheedge.com/reports/state-of-the-edge-report-2021

[5] https://www.pwc.de/de/technologie-medien-und-telekommunikation/5g-in-manufacturing.pdf

[6] Market and Technology Report "Artificial Intelligence Computing for Consumer", Yole Développement, 2019

[7] Market Report "Artificial Intelligence Computing for Automotive", Yole Développement, 2019

[8] Market and Technology Report "Artificial Intelligence for Medical Imaging", Yole Développement, 2020

[9] Market and Technology Report "Neuromorphic Sensing and Computing", Yole Développement, 2019

[10] https://www.vodafone.com/news/technology/multi-access-edge-computing-to-power-artificial-intelligence-for-automotive

[11] Armengaud E., Peischl B., Priller P., Veledar O. (2019) Automotive Meets ICT—Enabling the Shift of Value Creation Supported by European R&D. In: Langheim J. (eds) Electronic Components and Systems for Automotive Applications. Lecture Notes in Mobility. Springer, Cham. https://doi.org/10.1007/978-3-030-14156-1_4

[12] https://hadrianproject.eu

[13] https://teaching-h2020.eu

[14] FRACTAL project: https://fractal-project.eu

[15] ECSEL JU Integrated Development 4.0 project: http://www.idev40.eu

[16] TOWARDS A EUROPEAN DATA SHARING SPACE Enabling data exchange and unlocking AI potential, BDVA Position Paper, April 2019.

[17] A-Swarm project: https://www.tu-berlin.de/?209696

[18] ECSEL iRel4.0: https://www.irel40.eu

[19] ITEA3 COMPAS: https://itea3.org/project/compas.html

[20] http://www.evc1000.eu

[21] Samie F., Bauer L., Henkel J. (2019) Edge Computing for Smart Grid: An Overview on Architectures and Solutions. In: Siozios K., Anagnostos D., Soudris D., Kosmatopoulos E. (eds) IoT for Smart Grids. Power Systems. Springer, Cham. https://doi.org/10.1007/978-3-030-03640-9_2

[22] Z. Wang, M. Ogbodo, H. Huang, C. Qiu, M. Hisada and A. B. Abdallah, "AEBIS: AI-Enabled Blockchain-Based Electric Vehicle Integration System for Power Management in Smart Grid Platform," in IEEE Access, vol. 8, pp. 226409-226421, 2020, doi: 10.1109/ACCESS.2020.3044612.

[23] S. Chen et al., "Internet of Things Based Smart Grids Supported by Intelligent Edge Computing," in IEEE Access, vol. 7, pp. 74089-74102, 2019, doi: 10.1109/ACCESS.2019.2920488.

[24] Feng, Cheng & Wang, Yi & Chen, Qixin & Ding, Yi & Strbac, G. & Kang, Chongqing. (2020). Smart Grid Encounters Edge Computing: Opportunities and Applications. Advances in Applied Energy. 1. 10.1016/j.adapen.2020.100006.

[25] Huang, Yutao et al. "An Edge Computing Framework for Real-Time Monitoring in Smart Grid." 2018 IEEE International Conference on Industrial Internet (ICII) (2018): 99-108.

[26] Jennifer King and Christopher Perry: "Smart Buildings: Using Smart Technology to Save Energy in Existing Buildings", American Council for an Energy-Efficient Economy, Report A1701, 2017

[27] https://www.irel40.eu

[28] Michalos, G., Makris, S., Tsarouchi, P., Guasch, T., Kontovrakis, D., & Chryssolouris, G. (2015). Design Considerations for Safe Human-robot Collaborative Workplaces. Procedia CIRP, 37, 248–253. https://doi.org/10.1016/j.procir.2015.08.014

[29] Hadidi, R., Cao, J., Woodward, M., Ryoo, M. S., & Kim, H. (2018). Distributed Perception by Collaborative Robots. IEEE Robotics and Automation Letters, 3(4), 3709–3716. https://doi.org/10.1109/LRA.2018.2856261

[30] http://www.leti-cea.com/cea-tech/leti/english/Pages/Leti/Projects%20supported/Motion-project.aspx

[31] https://www.tyndall.ie/news/tyndall-welcomes--8-million-in-disruptive-technologies-innovation-funding

[32] European funding project Andante: https://cordis.europa.eu/project/id/876925/de

[33] M. Akhlaq M. S. Zareen, S. Tahir and B. Aslam. 2019. Artificial Intelligence/ Machine Learning in IoT for Authentication and Authorization of Edge Devices. 2019 International Conference on Applied and Engineering Mathematics (ICAEM) (2019), 220–224. https://doi.org/10.1109/ICAEM.2019.8853780

[34] https://www.eurotech.com/en/products/iot/iot-edge-framework/everyware-software-framework

[35] https://www.eurotech.com/en/products/iot

[36] OpenEI: An Open Framework for Edge Intelligence Xingzhou Zhang, Yifan Wang, Sidi Lu, Liangkai Liu, Lanyu Xu and Weisong Shi

[37] E. De Coninck et al. Distributed neural networks for Internet of Things: the Big-Little approach. Springer IoT360 2015: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp 484-492, vol. 170. 2016.

[38] S. Teerapittayanon et al. Distributed Deep Neural Networks over the Cloud, the Edge and End Devices. IEEE 37th International Conference on Distributed Computing Systems (ICDCS), pp. 328-339. 2017.

[39] Samek, Wojciech, et al., eds. Explainable AI: interpreting, explaining and visualizing deep learning. Vol. 11700. Springer Nature, 2019.

[40] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Van Der Driessche, J. Schrittwieser, I. Antonoglou and V. Panneershelvam, "Mastering The Game Of Go With Deep Neural Networks And Tree Search," Nature, 2016.

[41] A. Krizhevsky, I. Sutskever, G. Hinton, Ilya and E. Geoffrey, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84--90, 2017.

[42] G. Marcus, "Deep learning: A critical appraisal," arXiv preprint arXiv:1801.00631, 2018.

[43] F. Došilović, M. Brčić and N. Hlupić., "Explainable artificial intelligence: A survey," in 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2018, pp. 0210--0215.

[44] M. Masud, C. Woolam, J. Gao, L. Khan, J. Han, K. Hamlen and N. Oza, "Facing the reality of data stream classification: coping with scarcity of labeled data," Knowledge and information systems, vol. 33, no. 1, pp. 213--244, 2012.

[45] J. Vanschoren, "Meta-learning: A survey," arXiv preprint arXiv:1810.03548, 2018.

[46] T. Hospedales, A. Antoniou, P. Micaelli and A. Storkey, "Meta-learning in neural networks: A survey," arXiv preprint arXiv:2004.05439, 2020.

[47] A. Madela, A. Picon, C. Saratxaga, O. Belar, V. Cabezón, R. Cicchi, R. Bilbao and B. Glover, "Few shot learning in histopathological images: reducing the need of labeled data on biological datasets," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 1860--1864.

[48] Y. Wang, Q. Yao, J. T. Kwok and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," ACM Computing Surveys (CSUR), vol. 53, no. 3, pp. 1--34, 2020.

[49] D. Chen, Y. Liu, B. Kim, J. Xie, C. S. Hong and Z. Han, "Edge Computing Resources Reservation in Vehicular Networks: A Meta-Learning Approach," IEEE Transactions on Vehicular Technology, vol. 69, no. 5, pp. 5634--5646, 2020.

[50] Schoukens, J., & Ljung, L. (2019). Nonlinear System Identification: A User-Oriented Roadmap. ArXiv:1902.00683 [Cs]. http://arxiv.org/abs/1902.00683

[51] Laura von Rueden, Sebastian Mayer, Rafet Sifa, Christian Bauckhage and Jochen Garcke, "Combining Machine Learning and Simulation to a Hybrid Modelling approach: Current and Future Directions" IDA 2020, LNCS 12080, pp. 548–560, 2020.

[52] Von Rueden, Laura, et al. "Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning." Learning 18 (2019): 19-20.

[53] Urko Leturiondo, Oscar Salgado, Lorenzo Ciani, Diego Galar, Marcantonio Catelani, "Architecture for hybrid modelling and its application to diagnosis and prognosis with missing data," http://dx.doi.org/10.1016/j.measurement.2017.02.003

[54] Mei, B., Sun, L., & Shi, G. (2019). White-Black-Box Hybrid Model Identification Based on RM-RF for Ship Manoeuvring. IEEE Access, 7, 57691–57705. https://doi.org/10.1109/ACCESS.2019.2914120

[55] Glassey, Jarka, and Moritz Von Stosch, eds. Hybrid modelling in process industries. CRC Press, 2018. Link: https://ebookcentral.proquest.com/lib/bosch/reader.action?docID=5257196

[56] Hsi-Shou Wu, Energy-Efficient Neural Network Architectures, dissertation, University of Michigan, 2018

[57] Abhinav Goel, Caleb Tung, Yung-Hsiang Lu, and George K. Thiruvathukal, A Survey of Methods for Low-Power Deep Learning and Computer Vision, Purdue University and Loyola University, Chicago

[58] Yu Wang, Lixue Xia, Tianqi Tang, Boxun Li, Song Yao, Ming Cheng, Huazhong Yang, Low Power Convolutional Neural Networks on a Chip, Tsinghua University, Beijing, China, IEEE 2016

[59] Arianna Rubino, Melika Payvand, and Giacomo Indiveri, Ultra-Low Power Silicon Neuron Circuit for Extreme-Edge Neuromorphic Intelligence, Institute of Neuroinformatics, University of Zurich and ETH Zurich, IEEE 2019

[60] https://www.nature.com/scitable/blog/brain-metrics/are_there_really_as_many/

[61] https://aiimpacts.org/scale-of-the-human-brain/

[62] William B Levy, Victoria G. Calvert, Computation in the human cerebral cortex uses less than 0.2 watts yet this great expense is optimal when considering communication costs, doi: https://doi.org/10.1101/2020.04.23.057927

[63] Stolfa J. et al. (2020) Automotive Engineering Skills and Job Roles of the Future. In: Yilmaz M., Niemann J., Clarke P., Messnarz R. (eds) Systems, Software and Services Process Improvement. EuroSPI 2020. Communications in Computer and Information Science, vol 1251. Springer, Cham. https://doi.org/10.1007/978-3-030-56441-4_26

[64] Macher G., Druml N., Veledar O., Reckenzaun J. (2019) Safety and Security Aspects of Fail-Operational Urban Surround perceptION (FUSION). In: Papadopoulos Y., Aslansefat K., Katsaros P., Bozzano M. (eds) Model-Based Safety and Assessment. IMBSA 2019. Lecture Notes in Computer Science, vol 11842. Springer, Cham. https://doi.org/10.1007/978-3-030-32872-6_19

[65] Dimitrakopoulos, G., Uden, L., Varlamis, I.: The future of intelligent transport systems. Elsevier (2020)

[66] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

Authors

Bierzynski, Kay (Infineon)

Calvo Alonso, Daniel (Atos)

Gandhi, Kaustubh (Bosch Sensortec)

Lehment, Nicolas (NXP)

Mayer, Dirk (Fraunhofer ENAS)

Nackaerts, Axel (imec)

Neul, Reinhard (Bosch)

Peischl, Bernhard (AVL/AT)

Rix, Nigel (KTN)

Röhm, Horst (NXP)

Rzepka, Sven (Fraunhofer ENAS)

Seifert, Inessa (VDI/VDE-IT)

Steimetz, Elisabeth (VDI/VDE-IT)

Stree Bernard (CEA)

Tedesco, Salvatore (Tyndall)

Veledar, Omar (AVL/AT)

Wilsch, Benjamin (VDI/VDE-IT)

Imprint

Editing
Monika Curto Fuentes (VDI/VDE-IT) – Berlin, Germany

Graphic design / Layout
Juliane Lenz – Berlin, Germany

April 2021
smart-systems-integration.org

April 2021
smart-systems-integration.org