Edge AI: Reshaping the Future of Edge Computing with Artificial Intelligence

Conference Paper · October 2023

DOI: 10.5644/PI2023.209.07

2 authors:

Lejla Banjanovic-Mehmedovic
University of Tuzla
82 PUBLICATIONS **332** CITATIONS
SEE PROFILE

Anel Husaković
5 PUBLICATIONS **11** CITATIONS
SEE PROFILE

DOI: 10.5644/PI2023.209.07

Edge AI: Reshaping the Future of Edge Computing with Artificial Intelligence

Lejla Banjanović-Mehmedović[1*], Anel Husaković[2]

Abstract: *This paper highlights the growing importance of edge computing and the need for AI techniques to enable intelligent processing at the edge. Edge computing has emerged as a paradigm shift that brings data processing and storage closer to the source, minimizing the need for transmitting large volumes of data to remote locations. The integration of AI capabilities at the edge enables intelligent and real-time decision-making on resource-constrained devices. This paper discusses the significance of Edge AI across various domains, including automotive applications, smart homes, industrial IoT, and healthcare. By leveraging AI algorithms on edge devices, efficient implementation and deployment become possible, leading to improved latency, privacy, and security.The various AI techniques used in edge computing are presented, including machine learning, deep learning, reinforcement learning and transfer learning.*

As AI continues to play a pivotal role in driving edge computing, the integration of hardware accelerators and software platforms is gaining utmost significance to efficiently run inference models. A variety of popular options have emerged to accelerate AI at the edge, and notable among them are NVIDIA Jetson, Intel Movidius Myriad X, and Google Coral Edge TPU. The importance of specialized System-on-a-Chip (SoC) solutions for Edge AI, capable of supporting high-performance video, voice, and vision processing alongside integrated AI accelerators is presented as well.

By examining the transformative potential of Edge AI, this paper aims to inspire researchers, practitioners, and industry professionals to explore the vast possibilities of integrating AI at the edge. With Edge AI reshaping the future of edge computing, intelligent decision-making becomes seamlessly integrated into our daily lives, driving advancements across various sectors.

Keywords: *artificial intelligence, Edge AI, Edge computing, Edge Intelligence, embedded systems, FPGA, Industry 4.0, System-on-Chip (SoC)*

1. Introduction

Throughout the years, technology has undergone a series of transformations in data storage, transitioning from centralized mainframes to personal computing, and eventually to cloud computing. However, the current paradigm shift involves a departure from relying solely on data centers for hosting and

[*1]University of Tuzla, Faculty of Electrical Engineering, Tuzla, Bosnia and Herzegovina
[2]EaconZenica, Bosnia and Herzegovina
E-mail: lejla.banjanovic-mehmedovic@fet.ba

processing information. The concept of edge computing brings computation and data storage closer to the source, thereby minimizing the necessity of transmitting large volumes of data to remote locations. Edge computing refers to the paradigm of bringing computational power and data storage closer to the source of data generation, enabling faster processing and real-time decision-making. By processing data locally, edge computing reduces the time it takes for data to travel back and forth between the edge and the cloud, enabling faster response times and enhancing the reliability of applications. This approach reduces latency, conserves network bandwidth, enhances data privacy, and improves the overall efficiency of the system. It allows data to be filtered, aggregated, and analyzed locally, transmitting only the necessary insights or actionable information to the cloud[1-5].Edge computing concept is presented in Figure 1.

Edge computing allows for selective data processing at various levels. Subsequently, it categorizes the data into two main types [6]:

- ''Hot'' Data: Hot data refers to crucial signals that need to be transmitted to the production line's supervision system for immediate action.
- ''Cold'' Data: Cold data typically consists of historical data sets containing parameters such as pressure and temperature. This data is valuable for subsequent predictive analysis.

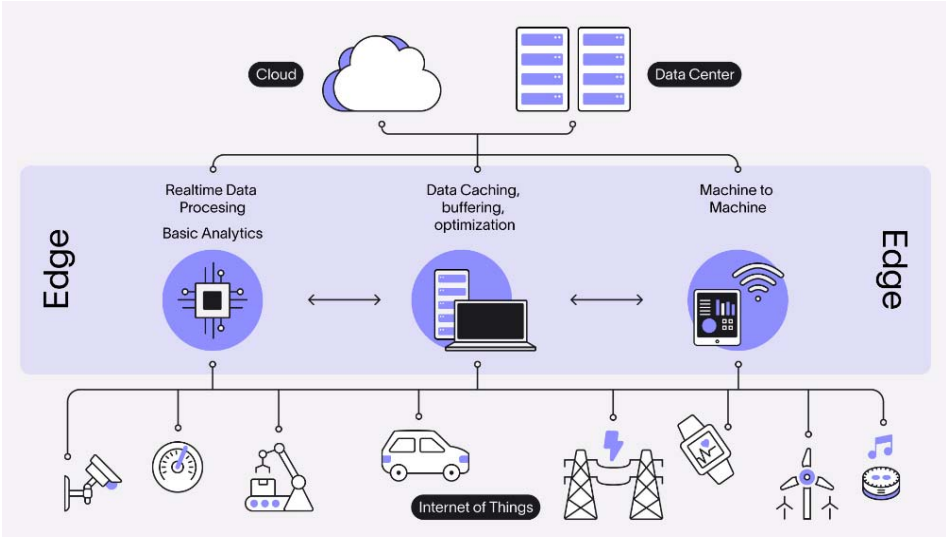


Figure 1. Edge Computing concept [7]

The edge computing applications cover a wide range of industries and use cases, fromsmart cities, autonomous cars, manufacturing, healthcare systems, toward Augmented Reality (AR) and Virtual Reality (VR), Figure 2.

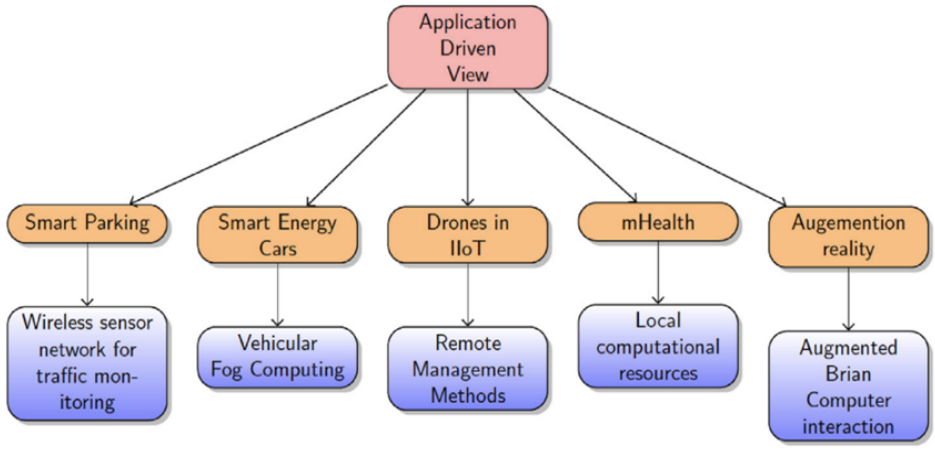


Figure2. Taxonomy of Edge computing applications [8]

In certain cases, there can be multiple layers, leading to the creation of the term ‘’fog computing’’ as an analogy to cloud computing.

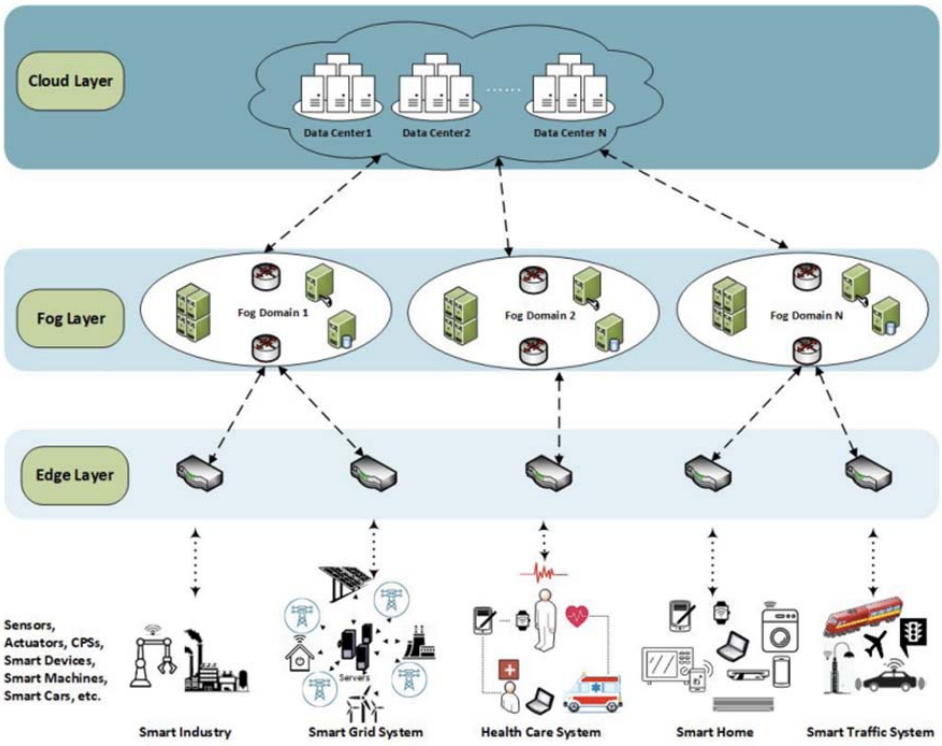


Figure 3. FogComputing concept [9]

Nowadays, the term ‘’fog computing’’ is gaining popularity within the industrial community. While ‘’cloud’’ signifies the highest and most widely distributed levels, ‘’fog’’ represents the intermediate levels situated between the edge and the cloud, as shown in Figure 3. It addresses some of the data congestion issues. At the same time, smarter devices are driving the data migration from cloud to edge. Edge AI refers to the deployment of artificial intelligence (AI) algorithms and models directly on edge devices, such as embedded systems, smartphones, Internet of Things (IoT) devices, and other local computing devices, rather than relying on cloud-based computing resources. In the Edge AI scenario, advanced AI models based on machinelearning (ML) algorithms will be optimized to run on the edge, as shown in Figure 4. It enables real-time complex data processing and analysis on the edge devices, reducing the need for constant data transmission to the cloud and enabling faster response times. Edge AI has several advantages over cloud-based AI solutions. Firstly, it reduces latency by processing data locally, which is crucial for applications that require real-time decision-making or low-latency responses. Secondly, it reduces the dependency on cloud connectivity, making it suitable for scenarios where there are limited or unreliable network connections.

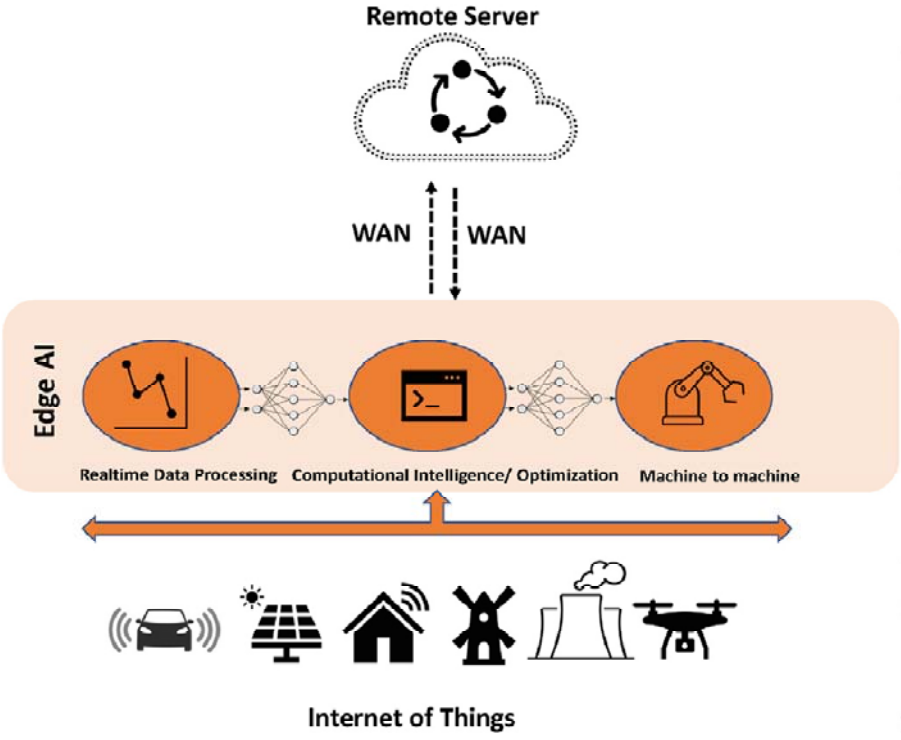


Figure 4.Edge AI concept[8]

Thirdly, it addresses privacy and security concerns by keeping sensitive data on local devices, minimizing the risk of data breaches, and ensuring data privacy.
Edge AI is increasingly being used in various domains, including smart homes, autonomous vehicles, healthcare, industrial automation, robotics, and surveillance systems. For example, in smart homes, edge AI can enable voice recognition and natural language processing directly on smart speakers or home automation hubs. In autonomous vehicles, edge AI enables real-time object detection and decision-making on-board the vehicle, reducing the dependence on cloud connectivity [10-11].
Edge Computing and Edge AI are related concepts but have distinct focuses at scope and objectives, computing infrastructure and functionality[1-2, 12-13]:

- *Scope and Objective.*Edge computing is a distributed computing paradigm that brings computing resources closer to the data source or the edge of the network. Its primary objective is to reduce latency, optimize network bandwidth, and enable real-time data processing and analysis at the network edge.Edge AI specifically refers to the deployment of artificial intelligence algorithms and models directly on edge devices. The objective of Edge AI is to enable AI capabilities at the edge, allowing local data processing, real-time decision-making, and intelligent behaviour without relying heavily on cloud resources.
- *Computing Infrastructure.*Edge computing involves the deployment of computing resources, such as edge servers, gateways, or routers, at the network edge. These devices provide computational power and storage capacity to process and analyze data closer to the source.Edge AI leverages the computing infrastructure of edge devices themselves, such as smartphones, IoT devices, or embedded systems. AI algorithms and models are deployed directly on these devices to enable local AI processing and decision-making.
- *Functionality.*Edge computing focuses on optimizing data flow, reducing latency, and improving the overall performance of applications by processing data closer to the edge. It facilitates tasks such as data aggregation, filtering, and forwarding to the cloud or other edge nodes.Edge AI builds upon edge computing by specifically incorporating AI capabilities at the edge. It allows for intelligent data processing, analysis, and decision-making on the edge devices themselves. Edge AI algorithms enable tasks like object recognition, predictive analytics, and real-time response directly on the edge devices.

2. Objectives and Functionalities of Edge AI

The key objectives of Edge AI are as follows:

- *Low Latency.* Edge AI aims to reduce the latency involved in data processing and decision-making by performing AI tasks directly on edge devices. This objective is important for real-time applications such as autonomous vehicles, industrial automation, and interactive systems[1].
- *Privacy and Security.* Edge AI focuses on keeping sensitive data on local devices, enhancing data privacy, and reducing the risk of data breaches. By processing data locally on edge devices, sensitive information can be kept within the local environment, reducing the risk of data breaches and unauthorized access. This objective is particularly important when dealing with sensitive data in applications like healthcare, finance, and surveillance systems [2,3].
- *Bandwidth Optimization.*Edge AI aims to optimize bandwidth consumption by minimizing the need for continuous data transmission to the cloud. By performing data processing and analysis at the edge, only relevant insights or summarized information can be transmitted, reducing the amount of data sent over the network. This optimization is crucial in scenarios with limited network connectivity or where bandwidth constraints exist[4].
- *Offline Operation.* Edge AI enables AI algorithms to operate effectively even in offline or intermittent connectivity scenarios. By deploying AI capabilities on edge devices, they can continue to function autonomously and make decisions locally without relying on constant cloud connectivity[2].
- *Energy Efficiency.* By processing data locally, Edge AI can reduce the amount of data sent over the network, leading to energy savings. Additionally, specialized edge AI hardware accelerators can further optimize energy consumption.
- *Scalability and Adaptability.*Edge AI aims to provide scalable and adaptable solutions that can cater to diverse edge devices and evolving application requirements. It focuses on developing lightweight and efficient AI algorithms that can run effectively on resource-constrained edge devices, ensuring flexibility and compatibility across different hardware platforms[14].
- *Real-time Data Analysis.* With Edge AI, data can be analyzed and filtered locally, enabling edge devices to provide real-time insights without the need for round-trip communication to the cloud.

The key functionality in Edge AI enables local data processing, real-time decision-making, contextual understanding, and resource optimization,

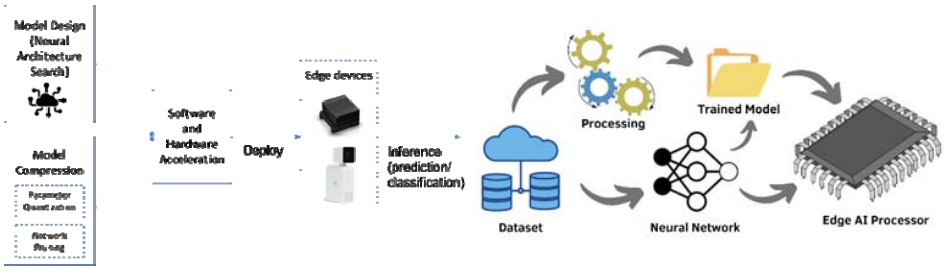
empowering edge devices to perform intelligent tasks and contribute to efficient and autonomous edge computing systems[2, 15-16]:

- *Data Processing and Analysis.* Edge AI algorithms are designed to process and analyze data generated by edge devices. This includes tasks such as data filtering, feature extraction, pattern recognition, and anomaly detection. By performing these computations locally, edge devices can extract meaningful insights from raw data and make informed decisions.
- *Real-time decision-making.* Edge AI enables edge devices to make autonomous and real-time decisions based on locally processed data. This includes tasks such as object recognition, classification, prediction, and control. By deploying AI models directly on edge devices, quick decision-making can be achieved, reducing the need for constant communication with the cloud.
- *Contextual Understanding.* Edge AI algorithms aim to understand the context in which the edge devices operate. This involves capturing and interpreting environmental cues, user interactions, and situational awareness. By understanding the context, edge devices can adapt their behaviour and responses accordingly.
- *Resource Optimization.*AI functionality in Edge AI also focuses on optimizing resource usage on edge devices. This includes techniques such as model compression, quantization, and optimization to reduce computational requirements and memory footprint. These optimizations ensure efficient execution of AI algorithms on resource-constrained edge devices.

3. The Key Technologies of Edge AI

Several key technologies collectively contribute to the efficient implementation and deployment of AI at the edge, enabling real-time and intelligent decision-making on resource-constrained devices:

- *Edge Data Management.* Efficient data management techniques, including data filtering, aggregation, and pre-processing, are employed to reduce the amount of data transmitted and processed at the edge. This helps in conserving network bandwidth and computational resources[17-18].
- *Model Optimization and Compression.*Model optimization techniques are employed to deploy AI models on edge devices with limited resources, as shown in Figure 5. These techniques include quantization, pruning, knowledge distillation, and model compression, reducing the model size and computational requirements while maintaining acceptable accuracy[19-20].



a). *b).*

Figure 5. Model Optimization and Compression: a). block-diagram,b). visual presentation [21,22]

- *Edge Device Hardware.*Edge AI accelerators are specialized hardware components designed to accelerate AI computations on edge devices, as shown in Figure 6. The advancements in hardware technologies, such as low-power processors, accelerators (e.g., GPUs, FPGAs), and dedicated AI chips (e.g., TPUs, NPUs), play a crucial role in enabling efficient execution of AI algorithms on edge devices with limited power and computational capabilities[23-24].

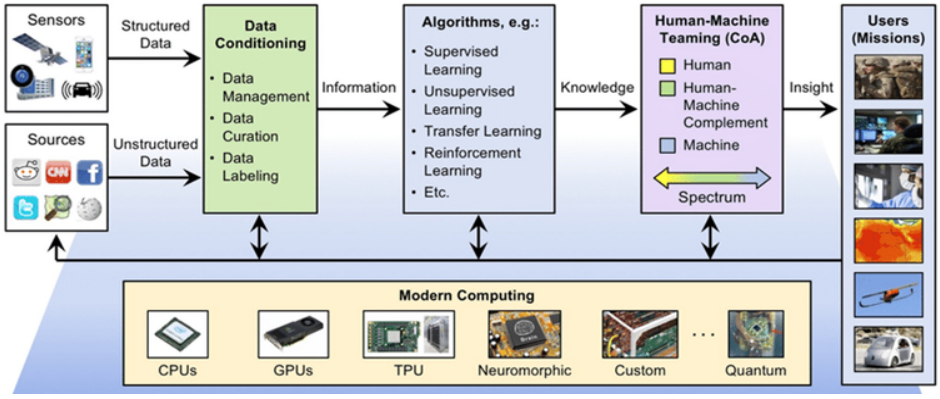


Figure 6. Edge Device hardware and AI algorithms in Edge AI [25]

- *Federated Learning.* Federated learning is a decentralized approach that allows multiple edge devices to collaboratively train a global AI model while keeping their data locally. The global model is updated with contributions from each device, minimizing data privacy concerns [26-27].

- *Edge-Cloud Collaboration.* In this paradigm, some parts of the AI processing occur on the edge device, while more resource-intensive or complex computations are offloaded to the cloud. This approach strikes a balance between local processing and cloud assistance to optimize performance and resource utilization. A real-time and robust fault detection approach through cooperation betweencloud and edge servers is presented in the paper [28].

3.1. Edge Device Hardware

FPGA (Field-Programmable Gate Array) and SoC (System-on-a-Chip) are two hardware technologies commonly used in Edge AI applications to accelerate AI computations and enable efficient deployment at the edge. FPGA is a reconfigurable hardware device that can be programmed to implement custom digital circuits and algorithms. In Edge AI, FPGAs are used to accelerate AI computations by offloading specific tasks or neural network models onto the FPGA hardware. FPGAs also offer flexibility as they can be dynamically reprogrammed or updated with new AI models or algorithms as needed[29-31].
SoC refers to a complete computing system integrated onto a single chip. It typically includes processing units (such as CPUs or GPUs), memory, I/O interfaces, and other components necessary for a functional system. SoCs provide a compact and power-efficient solution for deploying AI at the edge. They combine computational power with integrated components, enabling edge devices to perform AI tasks locally without relying on external computing resources. SoCs are commonly found in devices like smartphones, IoT devices, and edge servers, making them suitable for various Edge AI applications.
An architecture of on-device ML models is presented inFigure 7.

Figure7. Architecture of on-device ML models [32]

As AI continues to play a pivotal role in driving edge computing, the integration of hardware accelerators and software platforms is gaining utmost significance to efficiently run inference models. A variety of popular options have emerged to accelerate AI at the edge, and notable among them are NVIDIA Jetson, Intel Movidius Myriad X, and Google Coral Edge TPU. These cutting-edge solutions offer powerful capabilities to enable seamless AI processing at the edge of networks, catering to diverse applications and requirements[33].

1. *Vision Processing Unit (VPU).*Vision Processing Units (VPUs) are a critical component in efficiently handling demanding computer vision and edge computing AI tasks. They strike a balance between power efficiency and compute performance, making them ideal for such workloads. An example of a popular VPU is the Intel Neural Computing Stick 2 (NCS 2), which is built upon the Intel Movidius Myriad X VPU. The Myriad X VPU employs a clever architectural environment that minimizes data movement by running programmable computation

strategies alongside workload-specific hardware acceleration. A standout feature of the Myriad X VPU is the Neural Compute Engine, an intelligent hardware accelerator designed for deep neural network inference. It is fully programmable with the Intel Distribution of the OpenVINO Toolkit, allowing for flexible and efficient implementation of AI workloads.

With the Myriad Development Kit (MDK), developers can harness the power of the Myriad X VPU to create custom vision, imaging, and deep neural network applications using preloaded development tools, neural network frameworks, and APIs.

2. *Graphics Processing Unit (GPU).*The Graphics Processing Unit (GPU) is a specialized chip known for its rapid processing capabilities, particularly in handling computer graphics and image processing. In the context of AI at the Edge, GPUs play a crucial role in bringing accelerated performance in a power-efficient and compact form factor.One noteworthy device family enabling this accelerated AI performance at the Edge is NVIDIA Jetson. Take, for example, the NVIDIA Jetson Nano development board, which comes equipped with a 128-core GPU and Quad-core ARM CPU. Coupled with nano-optimizedKeras and TensorFlow libraries provided by the NVIDIA Jetpack SDK, it enables seamless execution of neural networks with minimal setup requirements.Not to be left behind, Intel has also entered the discrete graphics processor market with the release of the Xe GPUs. These GPUs, optimized for AI workloads and machine learning tasks, focus on achieving state-of-the-art performance with reduced power consumption.An example of architecture of NVIDIA Jetson Xavier NXA used for real-time deep lane detection system is presented in Figure 8.
3. *Tensor Processing Unit (TPU).* A Tensor Processing Unit (TPU) is a specialized AI hardware designed to efficiently execute machine learning algorithms, particularly those based on artificial neural networks (ANN). The Google Coral Edge TPU stands out as Google's purpose-built ASIC for edge AI processing.
   The Google Coral TPU, tailored for edge environments, serves as a powerful toolkit enabling local AI production. Its onboard device inference capabilities empower users to develop a wide range of on-device AI applications with several core advantages: very low power

consumption, cost-efficiency, and offline capabilities.Google Coral devices support various machine learning frameworks, including TensorFlow Lite, YOLO, and R-CNN, enabling tasks like Object Detection and Object Tracking in video streams from connected cameras. This makes them highly versatile for AI applications at the edge.

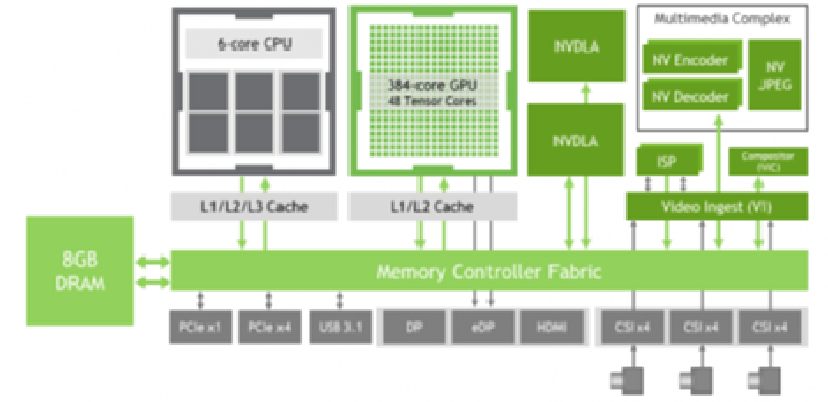


Figure 8. Architecture of NVIDIA Jetson Xavier NX[34]

Edge AI solutions have traditionally centered around camera vision, especially in automotive settings. However, in the context of smart homes, a more comprehensive approach is needed, incorporating video AI, voice AI, and vision AI for a multi-modal Human-Machine Interface (HMI). To achieve this, an ideal solution would involve a specialized System-on-a-Chip (SoC) specifically designed for smart homes, equipped to handle high-performance video, voice, and vision processing, while also featuring an integrated AI accelerator.

Such a solution would begin with an SoC platform that seamlessly integrates various types of processor engines, such as CPU, NPU, FPGA, and GPU, along with interfaces optimized for high-performance cameras and displays. This architecture enables a well-balanced combination of secure, cost-effective inferencing, and real-time, multi-mode performance.

The Synaptics Edge AI family offers a range of highly targeted SoCs, each tailored to address the specific demands of different consumer applications. Each SoC within this family comes equipped with the necessary processing cores and an appropriate level of integrated AI performance to best serve its intended application [35]. By providing a versatile and efficient edge AI solution, Synaptics aims to enhance the smart home experience and empower a seamless interaction between users and their connected devices.

One example of SoC device is the ''Versal AI Edge'' SoC that runs Linux on 1.76GHz Cortex-A72 with two cores and with two 750MHz Cortex-R5F cores, and additionally contains a flexible and FPGA-like ''adaptive core'' with up to 520K LUTs and an ''AI -ML'' core with up to 479 TOPS, as shown in Figure 9.

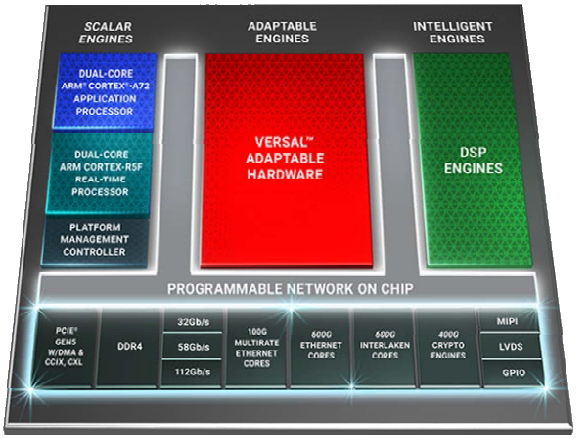


Figure 9. Architecture of Versal AI Edge[36]

Versal AI Edge SoC offers 10x higher computing power density compared to 16nm Zynq UltraScale+ and up to 4x AI performance per watt compared to Nvidia Jetson AGX Xavier module. The energy consumption of such systems starts with 6W, and for the needs of more demanding applications, they can be configured to work at up to 75 W.

4. AI techniques at Edge

Edge AI relies on a variety of artificial intelligence (AI) techniques to enable intelligent processing and decision-making at the edge devices themselves. Here are some common AI techniques used in edge AI [37-41], shown in Figure 10:

- *Machine Learning (ML).*Machine learning plays a crucial role in Edge AI applications by enabling edge devices to perform intelligent data analysis and decision-making locally. ML algorithms, including traditional techniques such as decision trees, K-Means and Fuzzy Clustering, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Multilayer feed-forward neural network[42], are employed in Edge AI to process data and extract meaningful insights. Neural networks are capable of learning complex patterns and extracting meaningful features from data.

- *Deep Learning (DL).*The most popular deep learning models in Edge AI applications are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to the input image, detecting features such as edges, corners, and textures. The pooling layers down sample the feature maps, reducing the spatial dimensions while preserving the important features. The fully connected layers at the end of the network perform classification or regression based on the learned features [43-44]. In edge AI applications, CNNs can be deployed on edge devices to perform tasks such as object detection, image classification, face recognition, and video surveillance. RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory) are types of neural network architectures commonly used in edge AI applications for sequential data analysis, such as speech recognition, natural language processing, and time series forecasting. They are well-suited for tasks that involve processing data with temporal dependencies and have the ability to retain information from past inputs[45-46].
- *Transfer Learning.* Transfer learning involves leveraging pre-trained models on large datasets and fine-tuning them on edge devices with limited data. This technique enables edge devices to benefit from the knowledge and features learned from powerful central servers or cloud-based models.

  Transfer learning is a powerful technique in Edge AI that allows leveraging pre-trained models on large-scale datasets and adapting them to specific tasks or domains with limited labeled data. By using transfer learning, edge devices can benefit from the knowledge learned from a source domain and apply it to a target domain, leading to improved performance and faster deployment [47-48].

- *Reinforcement Learning (RL).* RL techniques can be applied to edge AI scenarios where devices learn through trial and error to optimize their decision-making processes. RL enables autonomous learning and decision-making based on rewards and penalties received from the environment. RL (Reinforcement Learning) and DRL (Deep Reinforcement Learning) are powerful techniques used in edge AI applications to enable intelligent decision-making in dynamic environments [49-50]. DRL extends RL by incorporating deep neural networks to handle complex state and action spaces, enabling more

efficient and scalable learning. Edge AI, RL and DRL have several applications, such as autonomous systems, robotics, smart IoT devices, and resource management.

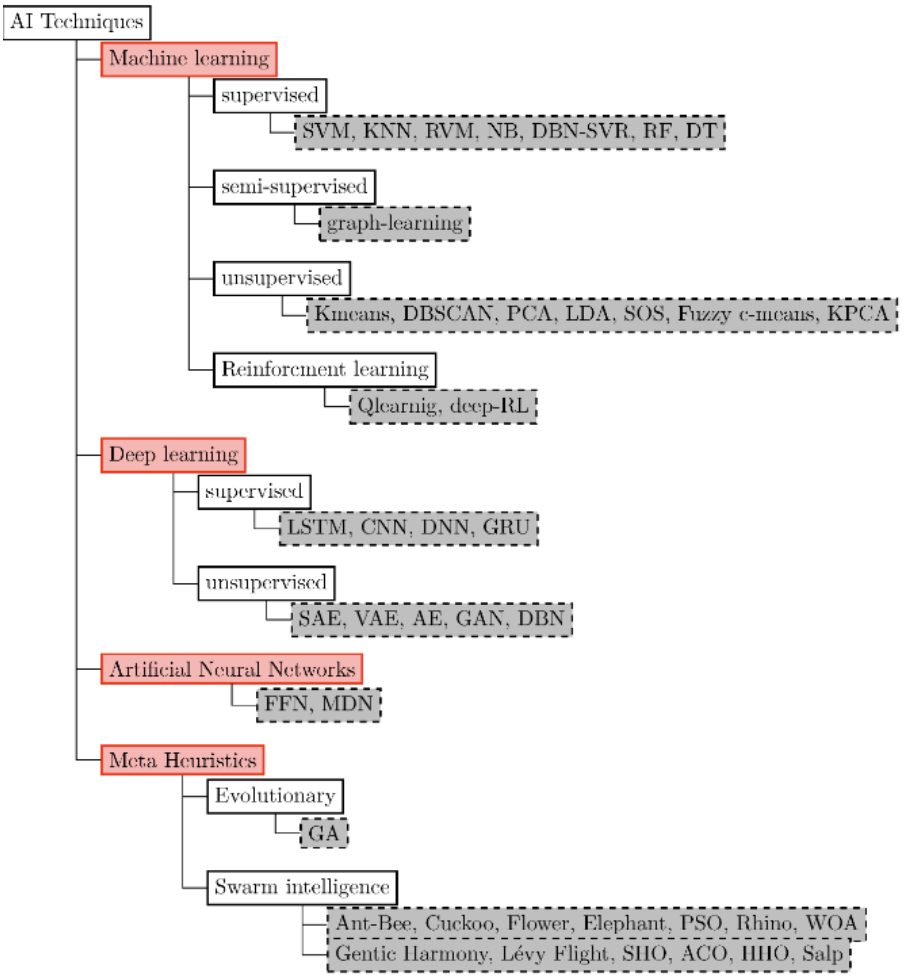


Figure 10. Categorization of AI techniques [41]

- *Meta Heuristics.*AI techniques are gaining increasing prominence in engineering as effective tools to tackle optimization problems [51]. Optimization tasks involve intelligent searches within large-dimensional spaces containing numerous decision variables, aiming to locate points that either minimize or maximize a specified objective function. Evolutionary computing, which draws inspiration from natural selection

and collective behavior patterns in nature, has emerged as a popular approach for optimization. Nature-inspired optimization heuristics, such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization Algorithm (ACO), Artificial Bee Colony (ABC), and Firefly algorithms, have demonstrated the ability to optimize diverse applications, surpassing the limitations of traditional optimization techniques.Among the subfields in this domain, two particularly relevant ones are genetic algorithms (GA) and swarm intelligence, exemplified by Particle Swarm Optimization (PSO). These AI-driven optimization techniques hold great promise in advancing engineering processes by efficiently solving complex optimization problems.

One example of fault detection using LSTM deep neural network at Edge is presented inFigure11.

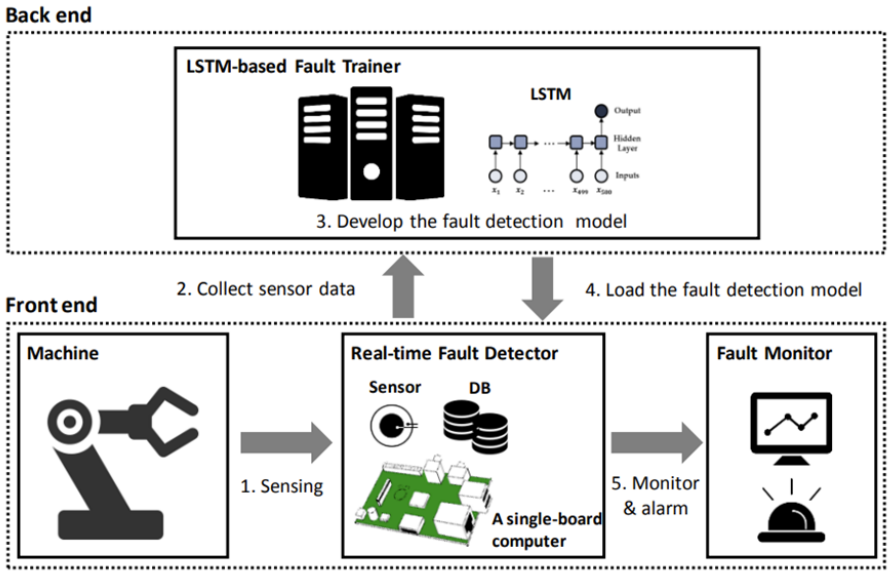


Figure 11. Example of fault detection using LSTM at Edge [52]

5. Edge AI Applications

The advancement of Edge Computing (EC) methodologies, encompassing robust IoT data, edge devices, storage, wireless communication, as well as security and privacy measures, has opened up opportunities for executing AI algorithms at the edge[53]. Edge Intelligence (EI) involves AI techniques such as machine learning, deep learning to enable intelligent data analysis and autonomous

decision-making at the edge. The diverse applications of EI, like real-time video surveillance, autonomous vehicles, industrial automation, smart healthcare systems, and Internet of Things (IoT) devices are presented in Figure 12.

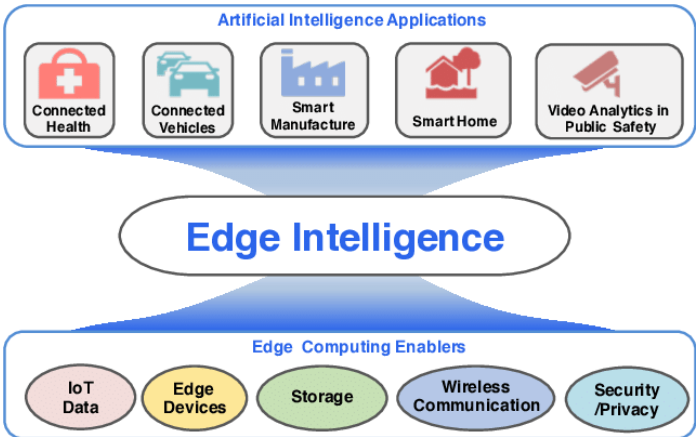


Figure 12. Edge Intelligence [53]

- *Autonomous Vehicles.* Edge AI is crucial for real-time decision-making in autonomous vehicles. It enables on-board perception, object detection, and collision avoidance by processing sensor data locally, ensuring quick response times and reducing dependency on cloud connectivity[54].
  An example of AI Edge solution for autonomous vehicle perception and control is presented at Figure 13. The partial end-to-end approach and end-to-end approach use deep neural network (DNN) for the final actuator output prediction [55].

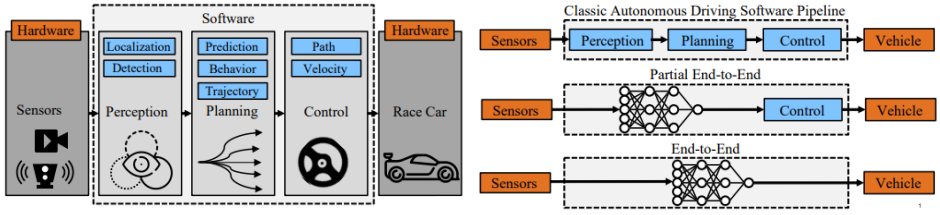


Figure 13. a). Example of autonomous vehicle AI Edge solution; b). Classic autonomous driving software pipeline in comparison to partialand full end-to-end software pipeline [55].

- *Smart Homes.* Edge AI allows smart home devices to perform voice recognition, natural language processing, and activity recognition locally. It enables faster response times, enhanced privacy, and local automation of various tasks within the smart home ecosystem [36].In tomorrow's smart

home, for example, loaded with artificial intelligence, the cameras could identify people and intrusion detection. Smart home Edge AI devices become faster, responding instantly to user commands and providing real-time information. These quicker response times could be life-saving like door locks with instant facial recognition or smart induction stoves that automatically change cooking temperature for smart home devices that contact emergency services or raise alarms [56]. Other examples may seem more futuristic. A refrigerator that can provide suggestions of what to make for dinner based on contents within the fridge. An oven that can tell you when your meal is cooked to perfection. A virtual personal home yoga trainer that can remind you to straighten your arms during a pose.

- *Industrial Automation.* Edge AI plays a vital role in industrial automation applications, such as predictive maintenance, quality control, and robotics. It enables real-time analysis of sensor data, anomaly detectionand autonomous decision-making at the edge, improving operational efficiency and reducing downtime[57]. One example of hydraulic system fault detection based on LSTM neural network via Edge Computing is presented in Figure 14.

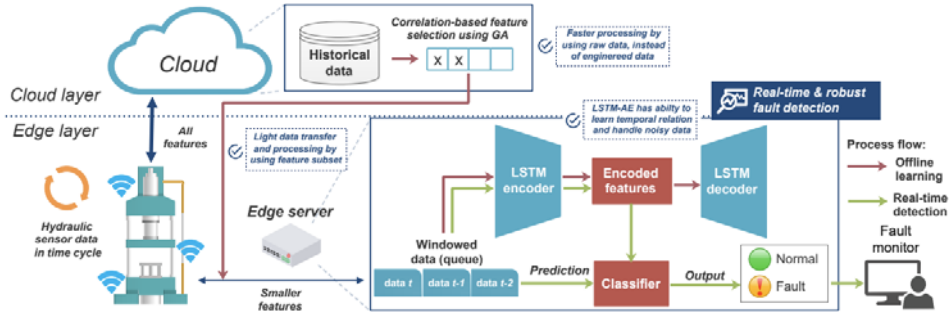


Figure 14. Example of Edge AI industry application [28]

- *Healthcare Monitoring.* Edge AI enables real-time monitoring and analysis of health data from wearable devices, such as heart rate monitors or glucose sensors, as shown in Figure 15. It allows for immediate detection of abnormal patterns, personalized healthcare recommendations, and timely alerts for medical interventions[58-59].

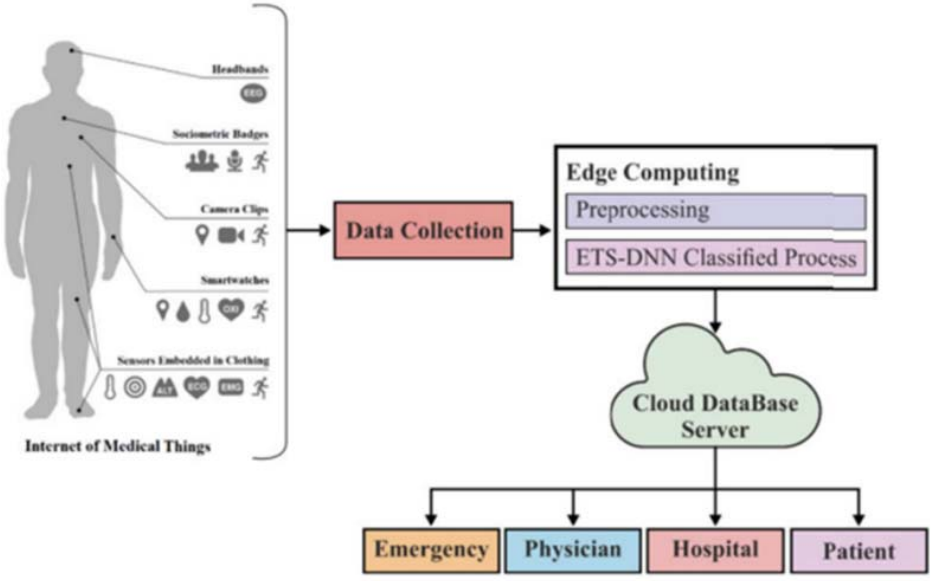


Figure 15. Smart healthcare with Edge AI solution [52]

- *Video Surveillance.* Edge AI is utilized in video surveillance systems to perform real-time object detection, tracking, and behavior analysis. It enables efficient video analytics at the edge, reducing bandwidth requirements and improving overall system performance[60]. An example of Edge AI camera application is presented in Figure 16.

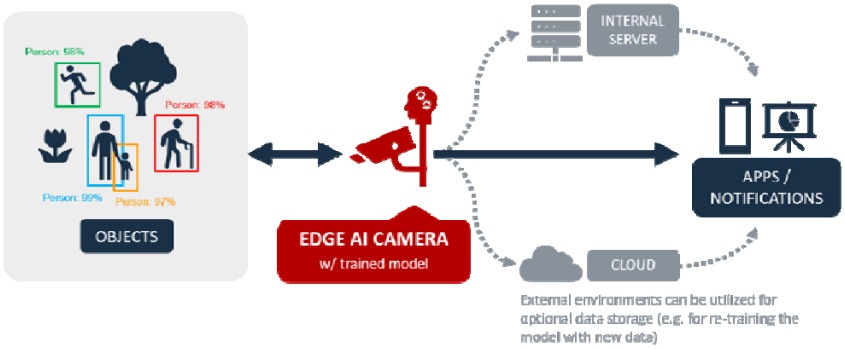


Figure 16. Example of Edge AI camera application [60]

- *Smart Cities.* Edge AI computing plays a crucial role in the development of smart cities. Sensors and cameras deployed throughout the city can collect data on traffic patterns, energy consumption, air quality, and more. Local edge servers can process this data to enable smart real-time traffic management, environmental monitoring, and resource optimization.A real-time deep lane detection system based on CNN Encoder-Decoder and Long Short-Term Memory (LSTM) networks for dynamic environments and complex road conditions is presented in [34].

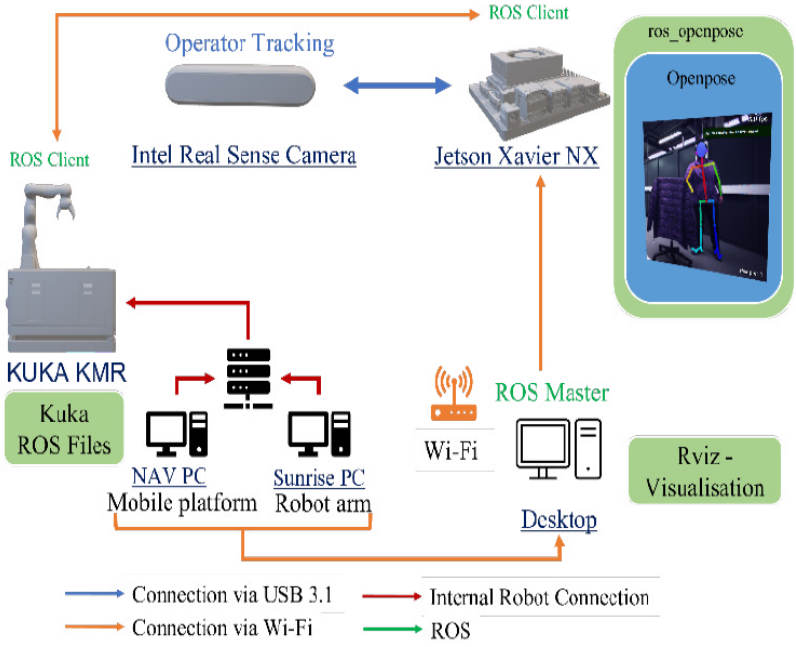


Figure 17. Overview of the integration of edge AI architecture for Collaborative Mobile Robots (CMR)[64]

- *Human-Robot Collaboration.* Edge AI plays a crucial role in enabling human-robot collaboration applications by bringing AI capabilities directly to the edge devices and robots involved in the collaboration. It allows robots to analyze and process data in real-time, make intelligent decisions, and interact with humans efficiently, without relying heavily on cloud resources. In human-robot collaboration scenarios, edge AI facilitates various aspects, such as perception, decision-making, and interaction. For example, edge AI algorithms deployed on robot platforms can enable real-time object recognition, tracking, and gesture recognition, enabling robots to understand and respond to human commands and actions. This enhances the safety and

efficiency of human-robot interactions[14, 61-63].The integration of edge AI architecture for Collaborative Mobile Robots (CMR) is presented in Figure 17.

The following table presents the most popular devices for edge AI applications.

Table 1. Comparison of various devices for Collaborative Mobile Robots (CMR) [64]

Device	GPU	CPU	Memory
Nvidia Jetson Xavier NX	384-core NVIDIA Volta	2-core (1900MHz at 15W)	8 GB LPDDR4x
Nvidia Jetson Nano	128-core NVIDIA Maxwell (472 GFLOPs)	4-core Ctx.A57 (1.43 GHz)	4 GB LPDDR4
Raspberry Pi 3B	None	4-core Ctx.A53 (1.2 GHz)	1 GB LPDDR2
Google Edge TPU	Vivante GC7000(32 GFLOPs)	Cortex-A53 (1.5 GHz main CPU)	1 GB LPDDR4

Furthermore, edge AI enables robots to make autonomous decisions based on local data processing. By deploying machine learning models directly on the robots, they can perform tasks such as task planning, navigation, and path optimization without continuous reliance on cloud connectivity. This leads to faster response times and increased autonomy in human-robot collaboration scenarios.

Additionally, edge AI provides data privacy and security benefits. Keeping sensitive data and AI algorithms on the edge devices ensures that personal information and critical decision-making processes remain within the local environment, minimizing the risk of data breaches and maintaining privacy.

The connection between Edge AI and Industry 4.0 and Industry 5.0 is a powerful synergy that holds immense potential for transforming the manufacturing landscape.

Edge AI plays a crucial role in enabling the realization of the Industry 4.0 vision by bringing AI capabilities to the edge of the network, closer to the industrial devices and machinery. One of the key advantages of Edge AI in Industry 4.0 is real-time data analysis and decision-making. By deploying AI algorithms directly on edge devices within the industrial environment, data can be processed and analyzed instantly, leading to faster response times and improved operational efficiency. This enables proactive maintenance, predictive analytics, and optimization of manufacturing processes, leading to reduced downtime, enhanced productivity, and cost savings. The edge devices like robots can perform intelligent tasks and make autonomous decisions without relying on a centralized cloud or network connection. This distributed intelligence allows for

greater flexibility, scalability, and resilience in the face of network disruptions, latency issues, or security concerns. Moreover, Edge AI enhances data privacy and security in Industry 4.0 applications. This is particularly crucial for industries dealing with proprietary designs, trade secrets, or compliance regulations[65-67].

The combination of Edge AI and Industry 4.0 also enables edge-to-cloud integration. While Edge AI empowers local decision-making and data processing, it can seamlessly connect with cloud-based platforms for broader analytics, machine learning model training, and centralized management. This hybrid approach leverages the strengths of both edge computing and cloud computing, creating a comprehensive and scalable architecture for Industry 4.0 applications.

Edge AI and Industry 5.0 are interconnected concepts that complement each other in the context of the evolving industrial landscape. Industry 5.0 emphasizes the collaboration between human workers and machines. Edge AI, with its ability to process data on the edge devices, facilitates seamless human-machine interaction by providing workers with contextual information and aiding them in complex tasks.Industry 5.0 creates more human-centric and sustainable production systems. Edge AI plays a critical role in enabling Industry 5.0 by bringing AI capabilities directly to the manufacturing floor. This allows workers to receive real-time feedback, make faster decisions, and respond to changing conditions promptly, thus leading to more efficient and flexible manufacturing systems[68].

The amalgamation of Edge AI and Edge Computing in the next-generation Industry 5.0, bolstered by 5G networks, presents significant opportunities across various sectors, including manufacturing, healthcare, Mobile Edge Computing (MEC), and smart cities. This technological advancement empowers industries to harness real-time intelligence, process data at the edge, and create intelligent, responsive systems that amalgamate the strengths of AI algorithms with the agility and low-latency capabilities of edge computing infrastructure[69-70].

6. Conclusion

Edge computing has gained significant traction in recent years due to its ability to process data closer to the source, reducing latency and improving efficiency.Edge AI techniques play a vital role in enabling intelligent decision-making at the edge devices themselves. The concepts of ‘’edge’’ and ‘’intelligence’’ are merely parts of a comprehensive solution that encompasses faster data processing, increased autonomy and transparency in operations, and a more agile and adaptable enterprise.

Edge AI and Industry 4.0 are intricately linked, with Edge AI serving as a catalyst for intelligent, decentralized, and autonomous systems in the

manufacturing domain. This integration brings real-time analytics, local processing, enhanced security, and edge-to-cloud connectivity, enabling smart factories, predictive maintenance, and optimized industrial operations. By leveraging the potential of Edge AI within Industry 4.0, organizations can unlock new levels of efficiency, productivity, and innovation in the manufacturing landscape.

The future of Edge AI is expected to be dynamic and transformative, as it continues to evolve and address various challenges. As edge computing becomes more prevalent, there will be continuous improvements in edge devices' hardware capabilities. This will lead to more powerful and energy-efficient processors and dedicated accelerators designed specifically for AI workloads.The demand for edge AI chips has experienced significant growth, and the market is projected to witness substantial expansion in the coming years. Edge AI and cloud computing will collaborate more seamlessly, creating a hybrid ecosystem where data and processing are dynamically distributed between edge devices and centralized cloud servers. This synergy will allow for more efficient and scalable AI applications.

7. References

[1] Satyanarayanan, M. (2017). The Emergence of Edge Computing.Computer, 50(1),30-39.doi:10.1109/MC.2017.9

[2] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing: Vision and Challenges. IEEE Internet of Things Journal, 3(5), 637-646. doi:10.1109/JIOT.2016.2579198

[3] Zhou, Z., Cao, J., Dong, J., Vasilakos, A. V., & Leung, V. C. M. (2019). Security and Privacy for Edge Computing: A Review. IEEE Internet of Things Journal, 6(1), 566-578. doi:10.1109/JIOT.2018.2846078

[4] Fernández-Caramés, T. M., & Fraga-Lamas, P. (2018). A Review on the Use of Edge Computing for IoT Security. Electronics, 7(11), 293. doi:10.3390/electronics7110293

[5] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing: The Cloud RAN Perspective. IEEE Communications Magazine, 54(12), 90-96. doi:10.1109/MCOM.2016.1600407CM

[6] https://www.forbes.com/sites/forbestechcouncil/2021/12/06/how-edge-computing-enables-industry-40/

[7] https://codeatelier.com/blog/edge-computing-rechenleistung-wo-sie-gebraucht-wird

[8] Raghubir Singh, Sukhpal Singh Gill, Edge AI: A survey, Internet of Things and Cyber-Physical Systems, Volume 3, 2023, pages 71-92, ISSN 2667-3452, https://doi.org/10.1016/j.iotcps.2023.02.004.

[9] https://www.insightsonindia.com/2019/10/30/edge-computing/Samy, Ahmed & Yu, Haining & Zhang, Hongli. (2020). Fog-Based Attack Detection Framework for Internet of Things Using Deep Learning. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2988854.
[10] Wang, Y., Li, P., & Zhang, C. (2019). A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. IEEE Internet of Things Journal, 4(5), 1125-1142. doi:10.1109/JIOT.2017.2787746
[11] Conoscenti, M., Vetro', A., &Catarinucci, L. (2018). Internet of Things: A survey on the security of IoT frameworks. Journal of Information Security and Applications, 38, 8-27. doi:10.1016/j.jisa.2017.11.006
[12] Zeng, Y., Zuo, C., & Sun, J. (2020). EdgeAI: A Survey on Edge Computing Technologies for Artificial Intelligence. IEEE Access, 8, 175204-175219. doi:10.1109/ACCESS.2020.3027920
[13] Tao, Y., Chen, M., Wang, Y., & Duan, Q. (2021). Federated Learning in Mobile Edge Computing: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, 23(4), 2181-2211. doi:10.1109/COMST.2021.3060884
[14] Gavrilovska, A., &Sardellitti, S. (2020). Edge Intelligence in IoT Era: Opportunities and Challenges. IEEEAccess, 8, 16422-16436. doi:10.1109/ACCESS.2020.2966724
[15] Chen, M., Zhang, Y., Hu, J., & Gonzalez, J. (2018). AIoT: When Artificial Intelligence Meets the Internet of Things. Mobile Networks and Applications, 23(2), 368-375. doi:10.1007/s11036-018-1092-2
[16] Mao, Y., You, C., Zhang, J., Huang, K., &Letaief, K. B. (2017). A Survey on Mobile Edge Computing: The Communication Perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322-2358. doi:10.1109/COMST.2017.2734284
[17] Hu, P., Patel, M., Sabaliauskas, A., Rundensteiner, E., & Ward, M. O. (2015). CEDAR: Centralized Data Access and Resource management in edge computing. In 2015 IEEE International Conference on Big Data (pp. 1627-1636).
[18] Zhou, X., Ma, H., & Deng, L. (2019). A Survey on Data Management for Internet of Things: A Edge Computing Perspective. IEEE Internet of Things Journal, 6(5), 8783-8800.
[19] Han, S., Mao, H., & Dally, W. J. (2015). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. arXiv preprint arXiv:1510.00149.
[20] Cheng, Y., Wang, D., Zhou, P., Zhang, T., & Zhang, H. (2018). Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey. Proceedings of the IEEE, 106(1), 21-43.
[21] https://darwinedge.com/edge-ai-deploying-ai-ml-on-devices/

[22] https://embeddedcomputing.com/technology/iot/edge-computing/edge-ai-is-overtaking-cloud-computing-for-deep-learning-applications

[23] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Herrmann, B. (2017). In-datacenter performance analysis of a tensor processing unit. In 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA) (pp. 1-12).

[24] Chen, T., Du, Z., Sun, N., Wang, J., Wu, C., Chen, Y., ... & Temam, O. (2014). DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In 2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA) (pp. 269-282).

[25] Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner, Survey of Machine Learning Accelerators, IEEE-HPEC conference, Waltham, MA, September 21-25, 2020.

[26] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1-19.

[27] Li, M., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 37(3), 50-60.

[28] Fawwaz, Dzaky Zakiyal, and Sang-Hwa Chung. 2020. ‘’Real-Time and Robust Hydraulic System Fault Detection via Edge Computing’’ *Applied Sciences* 10, no. 17: 5933. https://doi.org/10.3390/app10175933

[29] Chen, Y., Luo, X., Wu, Y., & Feng, C. (2021). FPGA Acceleration for Edge AI. ACM Transactions on Embedded Computing Systems, 20(1), 1-23. doi:10.1145/3452121

[30] Hamada, H., Wang, Z., &Hassanien, A. E. (2021). FPGA and Edge Computing for Deep Learning: Opportunities and Challenges. IEEE Access, 9, 31975-31985. doi:10.1109/access.2021.3057433

[31] Mao, Y., Jiang, H., Lin, S., Ma, Y., Zhang, W., & Guo, M. (2020). Survey on FPGA-Based Deep Learning Acceleration. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(6), 773-786. doi:10.1109/tetci.2019.2946005

[32] NazliTekin, Abbas Acar, Ahmet Aris, A. SelcukUluagac, VehbiCagriGungor,Energy consumption of on-device machine learning models for IoT intrusion detection,Internet of Things,Volume 21,2023,100670,ISSN 2542-6605,https://doi.org/10.1016/j.iot.2022.100670.

[33] https://viso.ai/edge-ai/ai-hardware-accelerators-overview/

[34] Yassin Kortli, SouhirGabsi, Lew F.C. Lew Yan Voon, Maher Jridi, MehrezMerzougui, Mohamed Atri,Deep embedded hybrid CNN–LSTM network for lane detection on NVIDIA Jetson Xavier NX,Knowledge-

Based Systems,Volume 240,2022,107941,ISSN 0950-7051,https://doi.org/10.1016/j.knosys.2021.107941.
[35] https://www.eetimes.com/edge-ai-solutions-for-smart-homes-can-transform-hmi/
[36] https://www.anandtech.com/show/16750/xilinx-expands-versal-ai-to-the-edge-helping-solve-the-silicon-shortage
[37] Li, S., Li, Y., Wang, C., So, H. C., & Li, B. (2020). Artificial intelligence at the edge: Deep learning-based edge computing for intelligent IoT systems. IEEE Signal Processing Magazine, 37(6), 22-30.
[38] Wang, Q., Wu, J., & Chen, J. (2020). Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing. Proceedings of the IEEE, 108(4), 450-467.
[39] Zhang, Y., Liu, Y., Chen, L., & Wu, D. (2020). An Overview of Artificial Intelligence Accelerators on Edge Devices. ACM Transactions on Embedded Computing Systems, 19(2), 1-23.
[40] Sharma, S., Mishra, D., Chung, T., & Kim, C. H. (2021). A Review on Edge AI: Machine Learning in Edge Computing Environment. IEEE Access, 9, 36398-36417.
[41] Bourechak, Amira, Ouarda Zedadra, Mohamed Nadjib Kouahla, Antonio Guerrieri, Hamid Seridi, and Giancarlo Fortino. 2023. "At the Confluence of Artificial Intelligence and Edge Computing in IoT-Based Applications: A Review and New Perspectives" *Sensors* 23, no. 3: 1639.
[42] Banjanović-Mehmedović, L. (2021). "Artificial Intelligence Advancement in Service Robots Applications", in Book: Service Robots: Advances in Research and Application (Editors: I. Karabegović, L. Banjanović-Mehmedović), Nova Science Publisher, USA.
[43] Gao, Y., Zhang, L., Sun, G., Cheng, M., & Liu, G. (2021). Edge Intelligence: On-Device Artificial Intelligence for Internet of Things. IEEE Communications Magazine, 59(2), 72-79.
[44] Zhang, Z., Ding, Z., &Jin, J. (2020). An Overview of Deep Learning in Edge Computing. IEEE Access, 8, 107042-107057.
[45] Li, Y., Liu, J., Wang, L., Yang, J., & Chen, J. (2019). Recurrent Neural Networks for Edge Computing: A Survey. IEEE Access, 7, 162184-162202.
[46] Zeng, S., Liu, X., Wang, K., & Chen, J. (2020). A Survey on Deep Learning for Edge Computing. IEEE Transactions on Industrial Informatics, 16(9), 6209-6221.
[47] Zhang, Z., Cao, L., & Xu, C. (2020). A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE, 108(4), 685-709.
[48] Yao, Q., Liao, H., Liu, S., Zhang, X., Zhang, J., & Mei, T. (2020). Efficient Transfer Learning with Meta Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[49] Luong, N. C., &Savkin, A. V. (2020). Reinforcement Learning in Edge Computing and IoT: Recent Advances and Open Challenges. IEEE Internet of Things Journal, 7(10), 9799-9812.
[50] Ye, Q., Sun, G., Li, G., & Zhang, X. (2019). Deep Reinforcement Learning for Edge Computing in IoT: Review, Open Issues, and Challenges. IEEE Internet of Things Journal, 6(2), 2270-2281.
[51] Lejla Banjanović-Mehmedović: "Artificial Intelligence Drives Advances in Human Robot Collaboration" in Book: Industrial Robots: Design, Applications and Technology (Editors: Isak Karabegović, Lejla Banjanović-Mehmedović), Nova Science Publisher, USA, 2020.
[52] Mainak Adhikari, AmbigavathiMunusamy,iCovidCare: Intelligent health monitoring framework for COVID-19 using ensemble random forest in edge networks,Internet of Things,Volume 14,2021,100385,ISSN 2542-6605,https://doi.org/10.1016/j.iot.2021.100385.
[53] Xingzhou Zhang, Yifan Wang, Sidi Lu, Liangkai Liu, LanyuXuandWeisong Shi (2019): OpenEI: An Open Framework for Edge Intelligence, 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS).
[54] Liang, M., Xu, B., Ren, C., & Cheng, L. (2020). A Survey on Edge Computing for Autonomous Vehicles: Technologies and Challenges. IEEE Transactions on Intelligent Transportation Systems, 21(9), 3784-3799. doi:10.1109/TITS.2019.2956153
[55] Johannes Betz, Hongrui Zheng, Alexander Liniger, Ugo Rosolia, Phillip Karle, Madhur Behl, Venkat Krovi, Rahul Mangharam: Autonomous Vehicles on the Edge: A Survey on Autonomous Vehicle Racing, IEEE Open Journal of Intelligent Transportation Systems, 2022.
[56] https://www.rcrwireless.com/20220613/network-infrastructure/why-ai-at-the-edge-is-the-future-of-smart-homes-reader-forum, Access: 10. July 2023.
[57] Sharma, P., & Kumar, N. (2020). Edge Intelligence for Industrial Internet of Things: A Comprehensive Review. IEEE Transactions on Industrial Informatics, 16(6), 4189-4206. doi:10.1109/TII.2019.2950968
[58] Fatima Alshehri, Ghulam Muhammad, A Comprehensive Survey of the Internet of Things (IoT) and Edge Computing in HealthcareDecember 2020IEEE Access PP(99):1-1DOI: 10.1109/ACCESS.2020.3047960
[59] Patil, S., &Gandomi, A. H. (2020). Edge Intelligence in Healthcare: Opportunities and Challenges. Artificial Intelligence in Medicine, 103, 101796. doi:10.1016/j.artmed.2020.101796
[60] Xu, Y., Cao, J., & Jiang, X. (2020). Edge AI in Smart Video Surveillance: Challenges, Advances, and Opportunities. IEEE Network, 34(5), 216-222. doi:10.1109/MNET.011.2000334

[61] Horváth, G., Barsi, Á., &Haidegger, T. (2019). A Survey on Human–Robot Collaboration in Industrial Scenarios. Robotics, 8(1), 2. doi:10.3390/robotics8010002
[62] Spina, G., Fumagalli, M., Riboldi, C., & Ferrigno, G. (2021). Edge Computing for Robot-Assisted Surgery: An Overview. Sensors, 21(12), 4017. doi:10.3390/s21124017
[63] Gupta, V., & Ranganathan, N. (2018). Towards Edge Intelligence: Characterizing Edge Analytics and its Applications. In 2018 39th IEEE Sarnoff Symposium (pp. 1-6). doi:10.23919/SARNOF.2018.8557352
[64] Aswin K Ramasubramanian, Robins Mathew, Inder Preet, Nikolaos Papakostas, Review and application of Edge AI solutions for mobile collaborative robotic platforms, Procedia CIRP, Volume 107, 2022, Pages 1083-1088, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2022.05.112.
[65] Hermenier, Fabien, et al. ‘’Edge Computing in the Industrial Internet of Things: A Survey.’’ IEEE Transactions on Industrial Informatics (2020).
[66] D'Angelo, G., Sallese, L., &Stracquadanio, G. (2020). Edge AI meets Industry 4.0: Challenges and perspectives. IEEE Transactions on Industrial Informatics, 16(8), 5186-5194. doi:10.1109/tii.2019.2956575
[67] Chen, C., Zhang, Y., Wu, Z., & Li, J. (2020). Industrial Internet of Things-Enabled Smart Manufacturing Systems: A Survey. IEEE Transactions on Industrial Informatics, 16(1), 3-17. doi:10.1109/tii.2019.2897587
[68] Jiang, Y., Liu, X., Gao, L., & Wang, L. (2021). Edge computing in industry 5.0: An overview, opportunities, and challenges. Future Generation Computer Systems, 119, 638-651.
[69] https://www.linkedin.com/pulse/industry-50-5g-role-edge-ai-computing-mintu-kumar-chetry/
[70] https://analyticsindiamag.com/edge-ai-and-its-transition-into-industry-5-0/