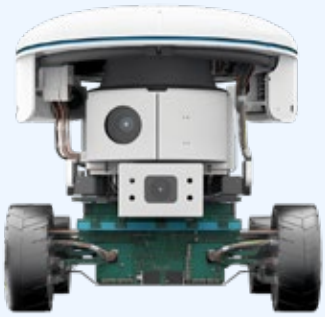
2023

EDGE AI TECHNOLOGY REPORT

The Guide to Understanding the State of the Art in Hardware & Software in Edge AI.






Introduction

The advent of Artificial Intelligence (AI) over recent years has truly revolutionized our industries and personal lives, offering unprecedented opportunities and capabilities. However, while cloud-based processing and cloud AI took off in the past decade, we have come to experience issues such as latency, bandwidth constraints, and security and privacy concerns, to name a few. That is where the emergence of Edge AI became extremely valuable and transformed the AI landscape.

Edge AI represents a paradigm shift in AI deployment, bringing computational power closer to the data source. It allows for on-device data processing and enables real-time, context-aware decision-making. Instead of relying on cloud-based processing, Edge AI utilizes edge devices such as sensors, cameras, smartphones, and other compact devices to perform AI computations on the device itself. Such an approach offers multitudes of advantages, including reduced latency, improved bandwidth efficiency, enhanced data privacy, and increased reliability in scenarios with limited or intermittent connectivity.

"Even with ubiquitous 5G, connectivity to the cloud isn't guaranteed, and bandwidth isn't assured in every case. The move to AIoT increasingly needs that intelligence and computational power at the edge."

Nandan Nayampally, CMO, Brainchip

While Cloud AI predominantly performs data processing and analysis in remote servers, Edge AI focuses on enabling AI capabilities directly on the devices. The key distinction here lies in the processing location and the nature of the data being processed. Cloud AI is suitable for processing-intensive applications that can tolerate latency, while Edge AI excels in time-sensitive scenarios where real-time processing is essential. By deploying AI models directly on edge devices, Edge AI minimizes the reliance on cloud connectivity, enabling localized decision-making and response.

The Edge encompasses the entire spectrum from data centers to IoT endpoints. This includes the data center edge, network edge,embedded edge, and on-prem edge, each with its own use cases. The compute requirements essentially determine where a particular application falls on the spectrum, ranging from data-center edge solutions to small sensors embedded in devices like automobile tires. Vibration-related applications would be positioned towards one end of the spectrum, often implemented on microcontrollers, while more complex video analysis tasks might be closer to the other end, sometimes on more powerful microprocessors.

"Applications are gradually moving towards the edge as these edge platforms enhance their compute power."

Ian Bratt, Fellow and Senior Director of Technology, Arm

When it comes to Edge AI, the focus is primarily on sensing systems. This includes camera-based systems, audio sensors, and applications like traffic monitoring in smart cities. Edge AI essentially functions as an extensive sensory system, continuously monitoring and interpreting events in the world. In an integrated-technology approach, the collected information can then be sent to the cloud for further processing.

Edge AI shines in applications where rapid decision-making and immediate response to time-sensitive data are required. For instance, in autonomous driving, Edge AI empowers vehicles to process sensor data onboard and make split-second decisions to ensure safe navigation. Similarly, in healthcare, Edge AI enables real-time patient monitoring, detecting anomalies, and facilitating immediate interventions. The ability to process and analyze data locally empowers healthcare professionals to deliver timely and life-saving interventions.

Edge AI application areas can be distinguished based on specific requirements such as power sensitivity, size limitations, weight constraints, and heat dissipation. Power sensitivity is a crucial consideration, as edge devices are often low-power devices used in smartphones, wearables, or Internet of Things (IoT) systems. AI models deployed on these devices must be optimized for efficient power consumption to preserve battery life and prolong operational duration.

Size limitations and weight constraints also play quite a significant role in distinguishing Edge AI application areas. Edge devices are typically compact and portable, making it essential for AI models to be lightweight and space-efficient. This consideration is particularly relevant upon integrating edge devices into drones, robotics, or wearable devices, where size and weight directly impact performance and usability.

Nevertheless, edge computing presents significant advantages that weren't achievable beforehand. Owning the data, for instance, provides a high level of security, as there is no need for the data to be sent to the cloud, thus mitigating the increasing cybersecurity risks. Edge computing also reduces latency and power usage due to less communication back and forth with the cloud, which is particularly important for constrained devices running on low power. And the advantages don't stop there, as we are seeing more and more interesting developments in real-time performance and decision-making, improved privacy control, and on-device learning, enabling intelligent devices to operate autonomously and adaptively without relying on constant cloud interaction.

"The recent surge in AI has been fueled by a harmonious interplay between cutting-edge algorithms and advanced hardware. As we move forward, the symbiosis of these two elements will become even more crucial, particularly for Edge AI."

Dr. Bram Verhoef, Head of Machine Learning at Axelera AI

Edge AI holds immense significance in the current and future technology landscape. With decentralized AI processing, improved responsiveness, enhanced privacy and security, cost-efficiency, scalability, and distributed computing, Edge AI is revolutionizing our world as we speak. And with the rapid developments happening constantly, it may be difficult to follow all the new advancements in the field.

That is why Wevolver has collaborated with several industry experts, researchers, professors, and leading companies to create a comprehensive report on the current state of Edge AI, exploring its history, cutting-edge applications, and future developments.

This report will provide you with practical and technical knowledge to help you understand and navigate the evolving landscape of Edge AI.
This report would not have been possible without the esteemed contributions and sponsorship of Alif Semiconductor, Arduino, Arm, Axelera AI, BrainChip, Edge Impulse, GreenWaves Technologies, Sparkfun, ST, and Synaptics. Their commitment to objectively sharing knowledge and insights to help inspire innovation and technological evolution aligns perfectly with what Wevolver does and the impact it aims to achieve.

As the world becomes increasingly connected and data-driven, Edge AI is emerging as a vital technology at the core of this transformation, and we hope this comprehensive report provides all the knowledge and inspiration you need to participate in this technological journey.

Samir Jaber
Editor-in-Chief

Chapter I: Overview of Industries and Application Use Cases

In recent years, there has been a clear shift of data from centralized cloud data centers to small-scale, local data centers and edge devices that reside close to the data sources. This has led to the emergence and rise of Edge AI. Specifically, the proliferation of data processed at or near the source of data generation has been a key enabler of Edge AI applications in different application sectors.

Nowadays, many enterprises are deploying and using edge functionalities as part of their AI applications. These functionalities enable them to develop energy-efficient, low-latency applications that exhibit real-time performance. Moreover, Edge AI functionalities provide enterprises with significant security and data protection benefits, which lead to improved privacy control and more effective compliance with applicable regulations. These benefits make Edge AI very appealing to organizations in many different sectors, which deploy and use Edge computing features in a variety of use cases.

This is the reason why the Edge AI market has a growing momentum. According to Fortune Business Insights, the Edge AI market is expected to grow from USD 15.60 billion in 2022 to USD 107.47 billion by 2029 at a Compound Annual Growth Rate (CAGR) of 31.7%.

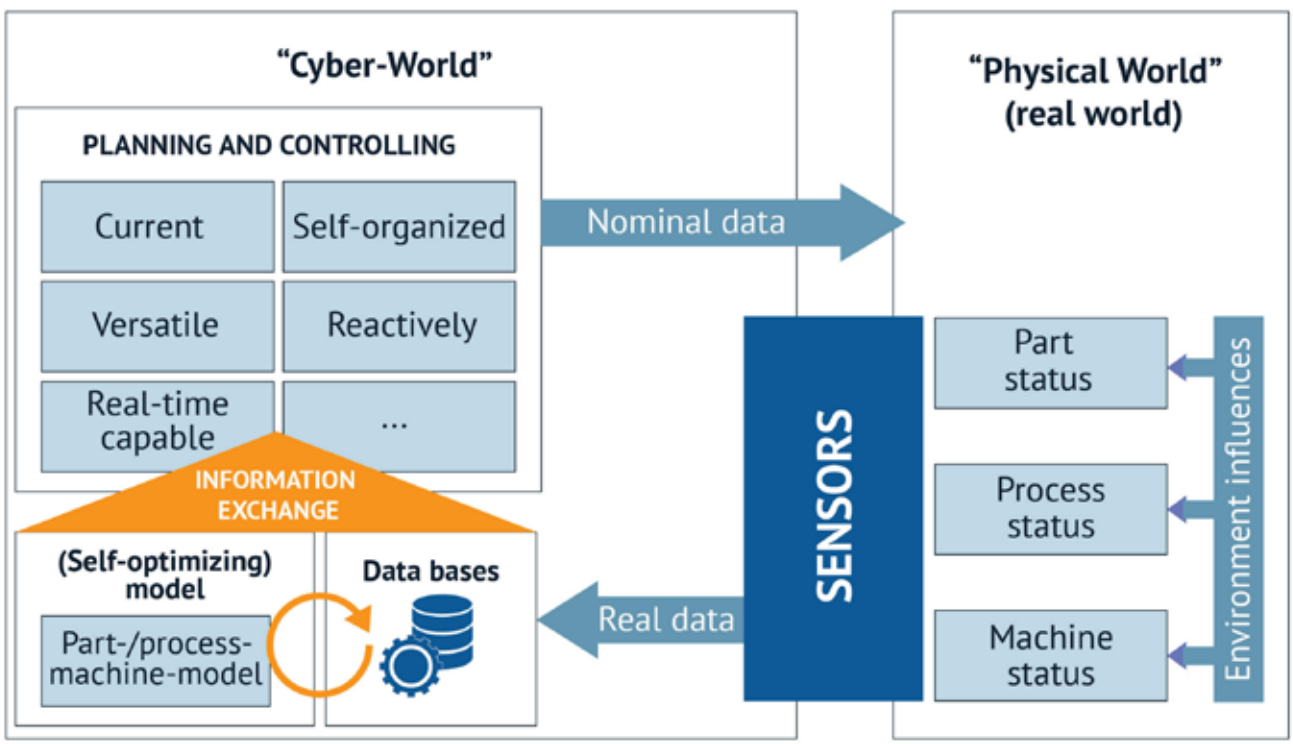
Industrial & Manufacturing

Innovators in the industrial sector see Edge AI and machine learning as vital technologies for their future business prospects. A survey fielded in the spring of 2023 by Arm found that edge computing and machine learning were among the top five technologies that will have the most impact in the coming years. In fact, nearly 70 percent of the respondents felt that IoT technologies were absolutely necessary for them to compete in their markets.

Industrial modernization and the shift to smart manufacturing have sparked innovations in automation, robotics, and industrial IoT (IIoT). The manufacturing sector has been undergoing a rapid digital transformation based on the introduction of Cyber-Physical Production Systems (CPPS) (e.g., industrial robots, intelligent automation devices) on the shop floor. These systems comprise a physical and a digital part, which enable the digitization of complex physical processes.

CPPS systems collect and analyze data about production processes such as production scheduling, quality inspection, and asset maintenance. Through data analysis, they derive unique insights about how to optimize these processes. Most importantly, they leverage these insights to close the loop to the manufacturing shop floor based on implementing real-time actuation and control functionalities. These functionalities significantly improve the efficiency and speed of automation tasks like product assembly and quality control.

Nevertheless, real-time actuation is hardly possible based on cloud data processing, which incurs significant latency. To alleviate the limitations of the cloud for real-time control, manufacturers are increasingly turning to Edge AI. This enables the execution of low-latency machine-learning functionalities on CPPSs, which makes them suitable for real-time actuation use cases.



Concept of a cyber-physical production system. Image credit: Imkamp, D. et al., J. Sens. Sens. Syst., 2016.

Some of the most prominent use cases of Edge AI in manufacturing include:

- Real-time detection of defects as part of quality inspection processes that leverage deep neural networks for analyzing product images

- Execution of real-time production assembly tasks based on low-latency operations of industrial robots

- Remote support of technicians on field tasks based on Augmented Reality (AR) and Mixed Reality (MR) devices; Low-latency Edge computing nodes are used to render AI-based AR/MR streams (e.g., AI-based repair recommendations) in real-time and effectively transfer on-the-job instructions from remote experts to on-site technicians.

While low latency is a primary Edge AI driver in the manufacturing sector, some use cases also benefit from Edge AI's security and privacy control features. For instance, several 3D printing use cases leverage Edge computing to avoid sharing sensitive Intellectual Property through a centralized cloud infrastructure.

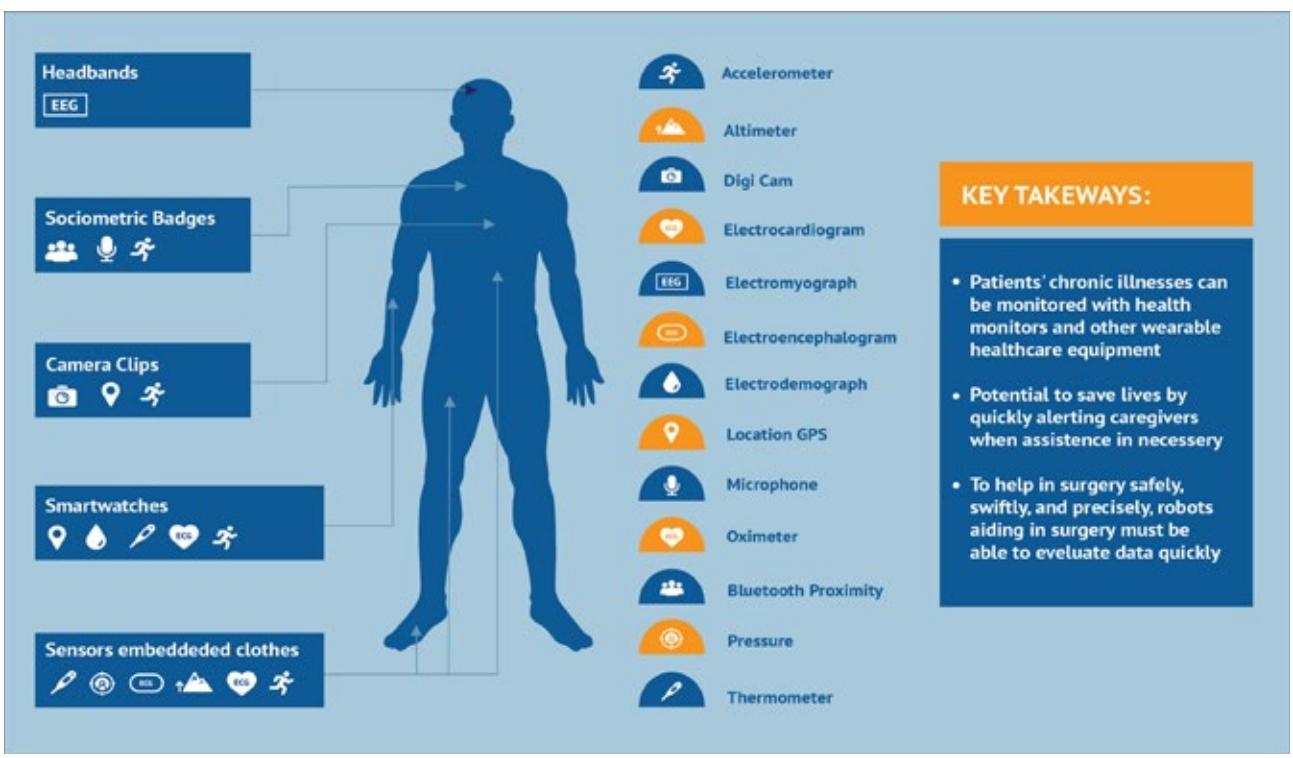
Healthcare

The healthcare sector produces and manages large amounts of sensitive data like patient datasets. Therefore, the rationale behind the deployment and use of Edge AI in healthcare lies in both the real-time functionalities and the privacy control features of this Edge paradigm. For instance, Edge AI deployments process data within the data owner's premises (e.g., care center, hospital, or patient's home), which is a compelling value proposition. Likewise, the low-latency characteristic of Edge AI enables real-time applications that can sometimes save a patient's life, especially with the deployment of wearable technologies in which machine learning is used to monitor vital signs, sleep patterns, and exercise levels. Some of the most prominent Edge AI applications in healthcare include:

- Remote patient monitoring and telemedicine, where healthcare professionals are provided with analytical insights about the patient's status without sharing patient data with other stakeholders
- Early diagnosis of disease indicators based on embedded machine learning on medical or wearable devices; Edge AI enables a host of novel diagnostic applications that deliver the merits of early screening and timely diagnosis at an unprecedented scale.
- Real-time surgery applications (e.g., telesurgery, robotic surgery), where low-latency Edge AI applications provide accurate guidance to the surgeon

The healthcare industry is currently preparing for the era of Edge AI through regulatory initiatives for AI-enabled medical devices, such as the European Medical Devices Regulation.

"One of the staggering statistics is the $1.1T loss in productivity for the US economy due to preventable chronic diseases," says BrainChip's Nandan Nayampally. "However, early diagnostics require constant monitoring and quite a bit of intelligent computation. To make these devices and services cost-effective and scalable, you need capable Edge AI."



Use cases of edge computing in healthcare.

Consumer Products

Edge AI brings several benefits to consumer products. First and foremost, it enables users' interactions with intelligent, low-latency, and high-performance AI-based interfaces, which improve the overall user experience. In this direction, Edge AI is used to implement perceptive interfaces based on speech, video, gestures, and other modalities.

Secondly, it can reduce latency when accessing data and services of consumer devices like smart appliances and gaming platforms. For example, by embedding AI functions in smart appliances, it is possible to implement functionalities for intelligent management, such as identifying the exact times when they need maintenance or repair.

Thirdly, Edge AI functionalities enhance the security and privacy of consumer products, as they reduce the amount of data shared outside the product's administrative domain. Finally, Edge AI functions improve consumer products with real-time data analytics capabilities. This is the case with gaming platforms that support real-time interaction functionalities with the end-user in the loop.

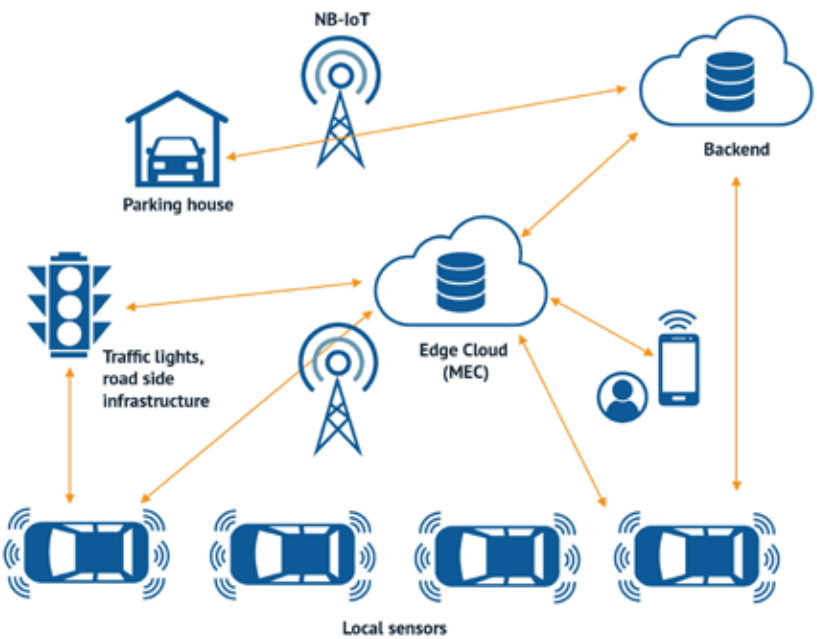
Consumer wearables, like the Oura Ring, use Edge AI to monitor health metrics and provide the user with real-time information, trends, and predictions. Image credit: Oura.

Transportation

The development of smart transportation systems that make decisions in real time is based on edge computing and Edge AI functionalities. Specifically, Edge AI functionalities enable a significant number of real-time decision-making applications such as connected driving applications, self-driving cars, automatic picking machines, and autonomous driving vehicles.

All these applications exhibit autonomous real-time behaviors implemented based on machine learning at the edge rather than based on data transfer and analytics within a cloud data center.

Self-driving cars and autonomous driving are sometimes considered the holy grail of Edge artificial intelligence. This is because autonomous vehicles deploy and use many Edge AI functions (e.g., deep neural networks at edge nodes) in order to see the road, understand the driving context, and make real-time decisions towards driving safely.



Architecture of autonomous vehicles with edge computing. Image credit: Jun-Ho H, Yeong-Seok S, IEEE Access, 2019. Adapted by Wevolver.

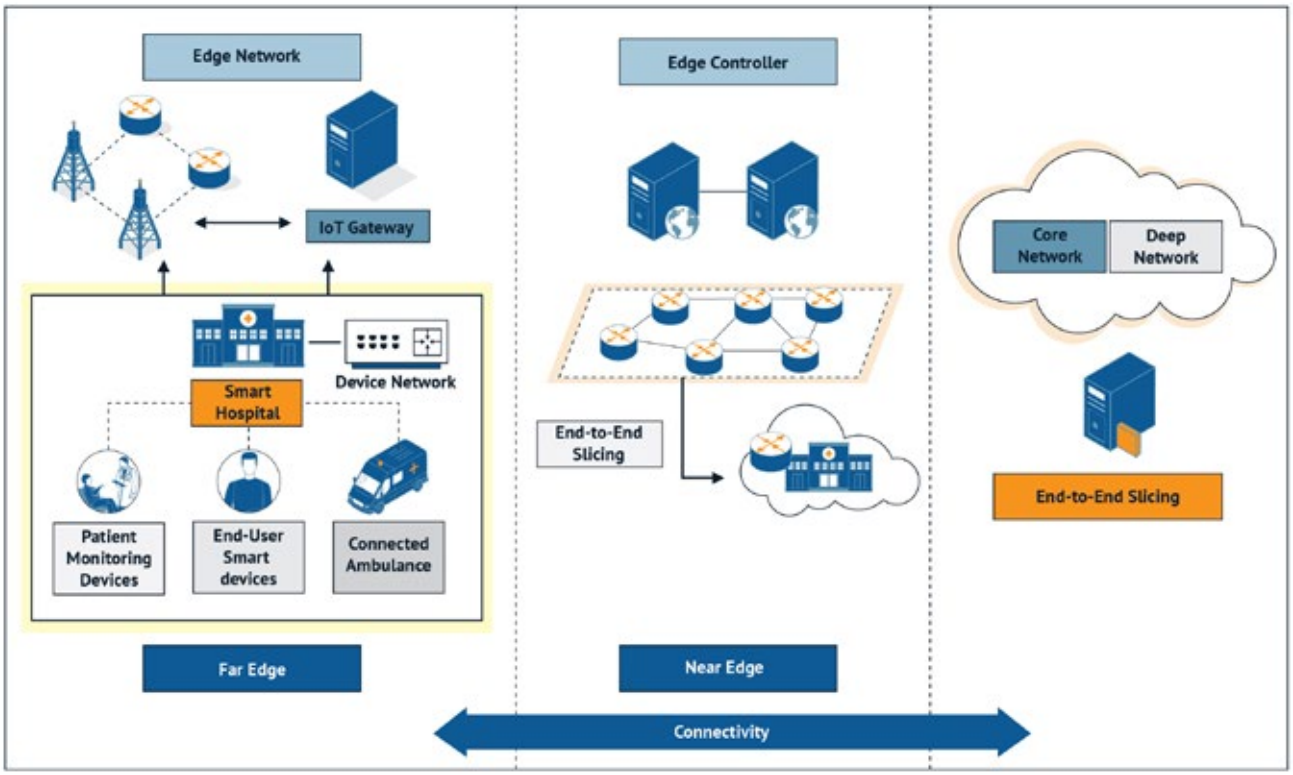
Smart Cities

Edge AI has a range of use cases in smart cities, notably use cases that process data within edge clusters or edge devices. As a prominent example, smart cities develop and deploy Edge AI infrastructures that support connected vehicles. These infrastructures process information at the edge by means of machine learning algorithms towards providing contextual, perceptive, and intelligent functionalities to vehicles.

Smart cities also leverage Edge AI to maintain utility infrastructures like electricity grids and water networks. For instance, Edge AI functions can identify smart meter faults and water leakages in real time. This helps to optimize the use of resources and minimize their depletion.

Edge AI in smart cities is also utilized to support a range of telecommunication use cases. For example, it is used to offer on-demand, machine-learning services to various industrial end users as part of private deployments. Such telecommunication services provide cost savings, low latency, and increased resilience, critical benefits for both municipal systems and citizens.

Another vital use for Edge AI is in healthcare monitoring systems in combination with IoT to improve healthcare facilities and infrastructure in smart cities. By implementing a large number of interconnected embedded sensors and IoT devices, such smart systems can provide important services to the community through data and connectivity. With the risk of chronic diseases and viruses that can spread dangerously fast, such as COVID-19, a robust healthcare infrastructure based on interconnected edge and IoT devices can allow for faster and earlier detection of diseases, thus facilitating well-timed treatment, reducing overall costs, and enabling the healthcare system of the city to keep up with or even prevent the viral spread.



An Edge AI-enabled IoT healthcare monitoring system for smart cities.
Image credit: Rathi, V. K. et al., Computers and Electrical Engineering, 2021.

Finally, one of the most popular smart city use cases for Edge AI is smart surveillance. Specifically, Edge AI solutions are deployed close to surveillance cameras to identify abnormal behaviors and other incidents of security interest as to enhance public safety and enforce traffic laws automatically. Rather than transmitting large volumes of video streams in the cloud, Edge AI applications transfer a much smaller number of streams following the detection of suspicious events or other abnormalities based on edge machine learning.

Smart Home

Edge AI is also popular in several smart home use cases, primarily in smart security. It involves the execution of machine learning functions within smart security devices like cameras, alarm systems, and video doorbells. These functions detect abnormalities and produce alarms in real time.

Another prominent use case in smart homes involves entertainment and infotainment applications. Specifically, Edge AI functions can be deployed to control smart home multimedia apps such as music and home cinema video. Such deployments are supported by smart appliances like smart TV sets. This can be seen with devices such as Alexa, which ushered in the era of ML in the home about nine years ago. It's a testament to how successful an application AI can be for the home that millions of them are equipped with edge AI technology, whether owners understand AI's role in the devices or not.

"From a segment perspective, the home is one of the most important areas where machine learning is being deployed."

Parag Beeraka, Senior Director, Segment Marketing, Arm's IoT business unit

Smart doorbells leveraging Edge AI Face Recognition technology to detect faces regardless of lighting and head position.
Image credit: Xailient.

Chapter II: Advantages of Edge AI

From Cloud to Edge AI

Following the introduction of cloud computing, there was a surge of interest in managing and analyzing data on the cloud. At that time, many enterprises considered the cloud a very appealing option for running data analytics and machine learning functions due to its scalability, capacity, and quality of service.

Over the years, however, many organizations voiced concerns against transferring and processing large volumes of data in the cloud. For instance, companies were reluctant to transfer sensitive corporate data to a remote data center they did not control. Others were concerned about the latency of several applications, as cloud data management and analytics incurred data transfer through a wide area network.

Motivated by the above-listed limitations of cloud data processing, many organizations have turned to edge computing to manage some of their data close to the data sources. Edge computing limits the amount of data to be transferred to the cloud for processing and analytics. Hence, it alleviates the latency, security, and privacy challenges of data analytics in the cloud.

Edge AI Advantages

Leveraging the merits of edge computing, Edge AI enables organizations to develop practical AI applications. These applications exhibit exceptional performance, low latency, robust security, improved privacy control, and power efficiency. Edge AI also enables organizations to make optimal use of network, computing, and energy resources, improving the AI applications' overall cost-effectiveness. Let's explore the advantages of Edge AI in more detail:

- **Reduced Latency:** Edge AI applications limit the amount of data transfers over wide area networks, as processing occurs close to the data sources rather than within the cloud. As such, data processing is faster, and the latency of the Edge AI application is reduced. Moreover, the transfer and execution of instructions from the AI applications to the field result in much lower latency. This is vital for several classes of low-latency AI applications, such as applications based on industrial robots and automated guided vehicles. Some other applications built around video in which the data had to be sent to the cloud are now being processed in the Edge not only because they can but because it's vital to assess what's going on in near real-time in situations like security cases.

- **Real-Time Performance:** The lower latency of Edge AI applications makes them suitable for implementing functionalities that require real-time performance. For instance, machine learning applications that detect events in real-time (e.g., defect detection in production lines, abnormal behavior detection in security applications) cannot tolerate delays associated with the transfer and processing of data from the edge to the cloud.

- **Enhanced Security and Data Protection:** Edge AI applications expose much fewer data outside the organizations that produce or own the data. This reduces their attack surface and minimizes the opportunities for malicious security attacks and data breaches. These are why Edge AI applications tend to be much more secure than their cloud counterparts.

- **Improved Privacy Control:** Many AI applications process sensitive data, such as data related to security, intellectual property, patients, and other forms of personal data. Edge AI deployments create a trusted data management environment for all these applications, providing more robust privacy control than conventional AI applications in the cloud. The reason is Edge AI applications limit the amount of data transferred or shared outside the organizations that produce or handle sensitive datasets.

- **Power-Efficiency:** Cloud data transfers and cloud data processing are extremely energy-savvy operations. Specifically, cloud I/O (Input/Output) functionalities are associated with significant carbon emissions. Most importantly, cloud AI is not green at all, as very large volumes of data are typically processed by GPUs (Graphical Processing Units) and TSUs (Tensor Processing Units). Edge AI alleviates the problematic environmental performance of cloud AI applications. It reduces the number of I/O operations and processes data within edge devices or edge data centers. Hence, it results in an improved overall CO2 footprint for AI applications.

- **Cost-Effectiveness:** Edge AI applications economize on network bandwidth and computing resources because they transmit and process much less data than cloud computing applications. Moreover, they consume less energy than cloud AI applications. As a result, Edge AI applications can be deployed and operated at a considerably lower cost than cloud AI deployments.

- **On-Device Learning:** Certain Edge AI applications can be executed within a single device, such as an IoT device or a microcontroller. This enables the development of powerful and intelligent devices, like System-on-Chip (SoC) devices. One of their main characteristics is that they can learn on the device, which is the foundation for giving machines intelligence capabilities that are hardly possibly based on cloud processing. BrainChip's Nandan Nayampally remarks, "With on-device learning, features extracted from trained models can be extended to customize model behavior without having to resort to extremely expensive model retraining on the cloud. This substantially reduces costs."

- **New Capabilities Enabling Novel Applications:** The integration of Edge AI systems within AI applications provides capabilities not available a few years ago. For instance, it enables a new class of real-time functionalities in transport, manufacturing, and other industrial settings. Hence, Edge AI is not only improving on cloud AI applications but also unlocking opportunities for innovative applications that were not previously possible.

In a conversation with Synaptics, we asked Shay Kamin Braun, Director of Product Marketing, about the most important advantages that Edge AI provides. He reiterated the advantages mentioned above and added some notable ones.

While "Edge AI provides a reduced latency in response time and a better UX without the network latency in the cloud," Kamin Braun also highlighted privacy and battery life benefits. In terms of privacy, "PII (personally identifiable information) stays on the device and doesn't go to the cloud. Only metadata goes to the cloud in most cases. This increases privacy," explained Kamin Braun. "As for power efficiency, running AI at the edge helps you reduce the amount of data transmitted to the cloud over the network, which saves power. As a result, you get longer battery life or energy star compliance."

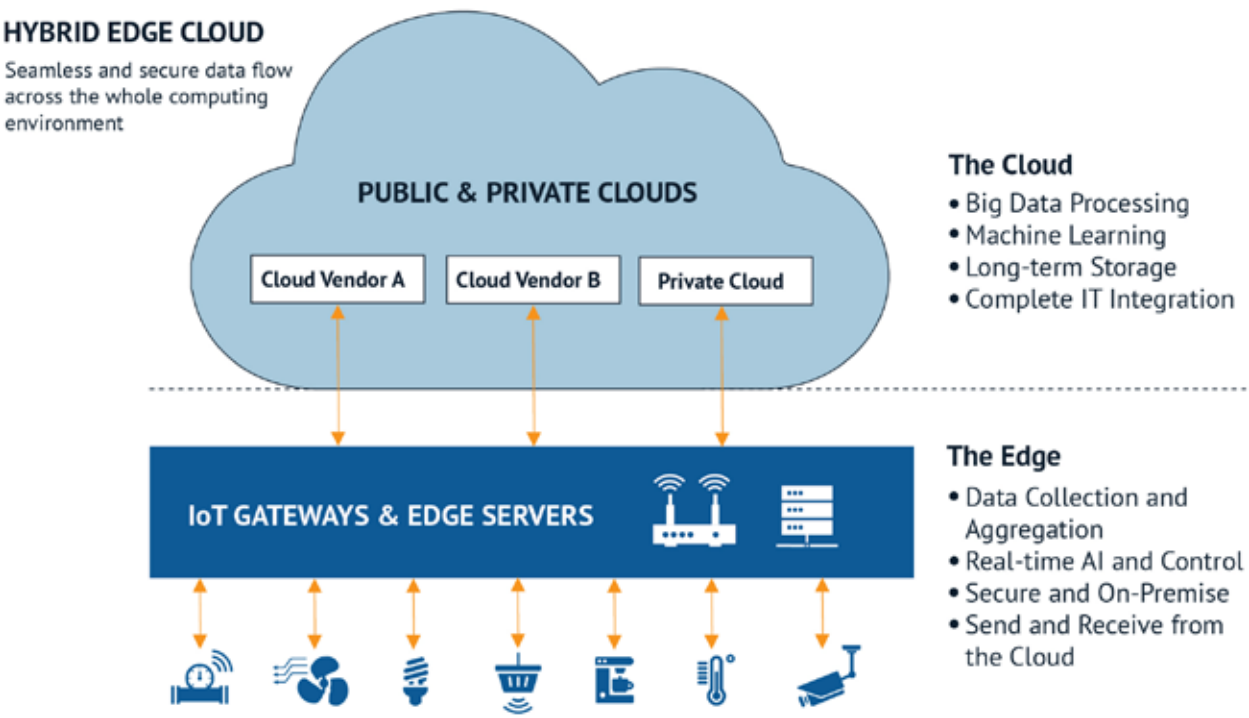
Kamin Braun also emphasized the advantage of availability, which can be an issue for Cloud AI given its inherent reliance on the network. "Maybe you have no coverage for mobile devices, or maybe the Wi-Fi is not working. You cannot always rely on the cloud for processing. On the contrary, if your application is completely at the edge, then you're not dependent on the network. Edge AI provides full availability."

"By doing the computation at the sensor with the context, you can optimize better than using a generic, aggregated solution. While the battery is a consideration, lower thermal footprint enables cheaper, cost-effective chip packaging and field use that really provides scale."

Nandan Nayampally, CMO, BrainChip

Cloud/Edge Computing Continuum: A Powerful Combination

With all the advantages that Edge AI has over cloud AI, many use cases still demand computationally intensive data processing within scalable cloud data centers. This is the case with deep learning applications, for example, which must be trained with large numbers of data points. In such cases, there is always a possibility to combine cloud AI and Edge AI paradigms to obtain the benefits of both worlds. Specifically, it is possible to perform a first round of local data processing at the edge to reduce the amount of uninteresting data and identify datasets that need processing in the cloud. Local data processing is usually followed by a transfer of selected data to the cloud, where data points from different edge data sources are aggregated and processed to extract global insights. Furthermore, cloud AI applications may push information about these insights back to the edge data sources, which enhances the intelligence of local processing and field actuation functions. This two-level data processing exploits different deployment configurations of edge devices in the cloud/edge computing continuum.



Hybrid solution of edge and cloud processing. Adapted from IEBMedia by Wevolver.

There are machine learning paradigms that inherently support this two-level data cloud/edge data processing approach. As a prominent example, federated machine learning employs local data processing to train machine-learning models across various edge computing nodes residing in different organizations. In this way, it produces a range of local machine-learning models. These are then fused in a central cloud location, which combines the local models into a more accurate global model.

This federated approach offers strong privacy and data protection characteristics, as source data need not leave the organizations where the edge nodes reside. Instead, an accurate global model is built in the cloud based on transferring and processing data about the local machine learning models. Therefore, the machine-learning models' information is shared while the source data are not.

The combination of cloud AI with Edge AI functionalities can be realized based on various deployment configurations, providing versatility at the levels of development, deployment, and operation. As a prominent example, in cases where local intelligence, real-time performance, and strong data protection are required, machine learning deployments can be restricted to the edge of the network. On the other hand, when the AI application involves the processing of high volumes of data (Big Data), Cloud AI is preferred.

Obviously, there are many use cases where AI processing can be distributed between cloud and edge. In such cases, the distribution should consider performance, power efficiency, and security trade-offs to yield the best possible configuration. Combining the merits of cloud computing and edge computing into a single cloud/edge AI deployment is yet another significant advantage of Edge AI systems.

Chapter III: Edge AI Platforms

Selecting and using the right Edge AI platform requires careful consideration of various factors. It is important to evaluate the project requirements, including latency, data processing capabilities, power consumption, size, weight, and heat dissipation. For example, if the project demands real-time processing and low latency, a platform with high processing capabilities and low latency is preferred. On the other hand, if power consumption is a concern, it is recommended to opt for an energy-efficient platform.

In addition to performance considerations, the level of expertise required to work with the platform should be taken into account. Some platforms are designed to be user-friendly, making them suitable for developers with varying levels of expertise. Conversely, other platforms may require specialized knowledge and skills.

The cost of the platform is also an important consideration. While some platforms are free, others may require licensing fees or additional costs for specific features. Economic feasibility should be evaluated based on the project's budget and resource constraints. Furthermore, the level of support and ecosystem compatibility the platform provides should be assessed. A platform with a robust support system and a thriving ecosystem can offer valuable resources, documentation, and community support, enabling developers to leverage the platform more effectively.

There is a wide array of Edge AI platforms available today, catering to the diverse needs of developers and enabling the deployment of machine learning models on edge devices. These platforms offer unique features and advantages instrumental in building cutting-edge AI applications. Among the prominent platforms are PyTorch Mobile, OpenVINO, NVIDIA Jetson, BrainChip Akida™, Caffe2, and MXNet. Each platform brings its own strengths and capabilities to the table, providing developers with a range of options. By carefully evaluating the project requirements and considering factors such as performance, expertise level, support, ecosystem, and cost, developers can make an informed decision about the most suitable platform for their specific Edge AI projects. In the subsequent sections, we will delve into each of these platforms, offering a brief overview and highlighting the key attributes that make them valuable tools for Edge AI development.

TensorFlow Lite

TensorFlow Lite is a robust and versatile platform designed for deploying machine learning models on Edge devices. It offers low latency, cross-platform compatibility, user-friendly features, and advanced capabilities, making it the preferred choice for developers. It enables quick and efficient inference on a wide range of platforms, such as mobile devices and microcontrollers. Through benchmarking tests, it has demonstrated impressive performance, achieving a median inference latency of 1.3 milliseconds on the Google Pixel 4 smartphone and 2.5 milliseconds on the Raspberry Pi 4 microcontroller board.

To further enhance its capabilities, TensorFlow Lite also supports hardware accelerators like the Google Coral Edge TPU, which effectively reduces latency. This platform empowers developers with diverse data processing capabilities, including image and audio recognition, natural language processing, and time-series analysis. It offers a collection of pre-trained models for these tasks and provides tools for training and optimizing custom models. TensorFlow Lite is adaptable to various input and output formats, such as image, audio, and sensor data, allowing it to cater to different use cases.

An important aspect of TensorFlow Lite is its accessibility. It is an open-source platform available for free, providing developers with a cost-effective solution. However, it is worth noting that using hardware accelerators like the Google Coral Edge TPU may incur additional expenses due to their higher cost.

When deploying deep learning models on mobile and edge devices, power consumption becomes a critical consideration. TensorFlow Lite addresses this concern by being lightweight and optimized specifically for these devices. It has a small memory footprint and can seamlessly integrate into mobile apps and microcontroller projects. In addition, it pays attention to heat dissipation, a crucial factor for devices with size and power constraints. By prioritizing efficiency and lightweight design, TensorFlow Lite minimizes heat generation, ensuring optimal performance.

The platform offers a comprehensive range of development tools and APIs, enabling developers to train and deploy their models across various platforms. From a technical perspective, TensorFlow Lite is designed to be flexible and extensible. Developers can customize and fine-tune their machine-learning models to suit their requirements. Furthermore, TensorFlow Lite supports multiple programming languages, including Python, C++, and Java, enhancing its versatility and ease of use.

The Akida edge neural network processor from BrainChip is also supported under TensorFlow using MetaTF, which adds functions to optimize the network for execution on the high-performance Akida hardware. As Peter van der Made, CTO and founder at BrainChip, explains, "One outstanding feature of Akida is that it executes the entire network internally without dependence on the CPU's resources, thereby significantly increasing computing power at the edge." Its event-based processing method is fast and power-efficient, running most networks within the microwatt or milliwatt range. The on-chip learning method allows for adding additional functions after the device is trained.

PyTorch Mobile

PyTorch is a widely adopted open-source machine learning framework that has recently extended its capabilities to include mobile devices through PyTorch Mobile. Its ease of use, flexibility, and support for state-of-the-art models make it a compelling choice for developers. PyTorch Mobile is optimized for speed and efficiency, offering low latency that enables its deployment on resource-constrained, low-power devices.

According to PyTorch, models built using this platform can achieve latencies of less than ten milliseconds on select mobile devices, which is sufficiently fast for many real-time applications. However, the actual latency of a PyTorch Mobile model depends on various factors, including the hardware platform, model complexity, and input data size.

PyTorch Mobile empowers developers with a broad range of data processing capabilities, including image recognition and natural language processing. The costs associated with using PyTorch Mobile vary depending on the hardware platform, model size, complexity, and specific use case requirements.

Power consumption is crucial when implementing machine learning on edge devices, mainly due to the limited battery life in many such devices. PyTorch Mobile addresses this concern with a small memory footprint, enabling efficient execution on low-power devices with restricted resources.

Additionally, PyTorch Mobile supports various hardware platforms, such as CPUs, GPUs, and DSPs, enabling developers to optimize their machine-learning models for specific hardware architectures and minimize power consumption.

Designed for lightweight and efficient operation, PyTorch Mobile has a compact binary size of approximately 4 MB, making it well-suited for mobile devices. It can also run on diverse hardware platforms, including microcontrollers with limited storage and processing capabilities, thereby catering to small form factors and weight constraints. The amount of heat dissipation experienced will depend on the specific hardware platform employed for deployment, as well as the size and complexity of the model.

Finally, PyTorch Mobile provides optimized models and tools for efficient and low-power inference, helping reduce the heat generated by the platform. It supports a variety of frameworks and programming languages, such as Python, offering adaptability to different use cases.

OpenVINO

OpenVINO (Open Visual Inference and Neural Network Optimization) is an agile platform that offers numerous benefits for deploying deep learning models. Its exceptional performance, support for multiple hardware platforms, and flexibility make it a preferred choice. Developed by Intel, OpenVINO enables efficient deployment across diverse platforms, including CPUs, GPUs, and FPGAs. Based on benchmarking tests conducted by Intel, OpenVINO achieves inference latencies as low as five milliseconds on specific hardware configurations, making it suitable for real-time applications like object detection and image recognition.

The platform provides optimized libraries and tools for popular deep learning frameworks such as TensorFlow, PyTorch, and Caffe, empowering developers to train and deploy models on various platforms. The costs of using OpenVINO depend on factors like model size, complexity, and the hardware platform employed.

In terms of power consumption, OpenVINO delivers optimized inference performance, resulting in lower power consumption than traditional CPU-based solutions. It also supports hardware accelerators like FPGAs, known for their low power consumption. OpenVINO has a relatively small footprint, and its size varies based on the selected hardware and software components for deployment. The platform's support for hardware accelerators reduces heat dissipation compared to traditional CPU-based solutions.

OpenVINO offers seamless integration with multiple programming languages, including C++, Python, and Java. It also integrates with other Intel tools and technologies, such as the Intel Distribution of OpenVINO Toolkit and the Intel System Studio.

Finally, OpenVINO achieves high performance through a combination of model optimization techniques like model quantization and compression. These techniques reduce the size and complexity of deep learning models, resulting in enhanced efficiency and faster execution on edge devices.

NVIDIA Jetson

NVIDIA Jetson offers a multitude of advantages, including high processing power, low latency, and support for various neural networks and development tools. One of the primary strengths of NVIDIA Jetson for edge AI lies in its exceptional processing power. The Jetson devices are equipped with NVIDIA GPUs specifically designed for high-performance computing in deep learning applications. These GPUs are optimized for parallel computing, enabling them to handle substantial data volumes within short timeframes efficiently. This makes Jetson devices particularly well-suited for real-time data processing requirements in applications like video analytics and autonomous robotics.

Another notable benefit of NVIDIA Jetson is its low latency. These devices feature high-speed interfaces that facilitate swift communication between the processor and other components. As a result, data can be processed rapidly, minimizing the time needed for decision-making in edge AI applications. Furthermore, Jetson devices find utility across a broad spectrum of edge AI use cases, encompassing smart cameras, drones, and industrial automation.

It is worth noting that NVIDIA Jetson devices exhibit relatively higher power consumption due to their significantly enhanced data processing capabilities. Consequently, additional cooling measures are often necessary to manage the generated heat effectively. NVIDIA Jetson provides a diverse selection of neural networks and development tools. This ecosystem empowers developers with an extensive array of options to facilitate their edge AI projects effectively.

Regarding costs, the price of using NVIDIA Jetson devices can vary depending on the chosen model and specifications. Nonetheless, these devices are designed to be cost-effective solutions for edge AI applications. For instance, the NVIDIA Jetson Nano Developer Kit combines high performance with a competitive price, enhancing accessibility for developers.

Additionally, NVIDIA Jetson devices are compact and lightweight, making them ideal for deployment in space-constrained environments where size and weight considerations are crucial factors to address.

Edge Impulse

Edge Impulse is an end-to-end platform designed to create, train, and deploy machine learning models on edge hardware. Compared to other Edge AI

platforms, Edge Impulse offers distinct advantages in terms of ease of use, flexibility, and device compatibility.

One key advantage of Edge Impulse is its user-friendly nature. The platform is designed to cater to users with varying levels of experience in machine learning and embedded systems. With its intuitive web-based interface, users can effortlessly create, train, and deploy machine learning models without requiring specialized hardware or software. This simplicity makes Edge Impulse an appealing option for enterprises and developers who seek quick prototyping and testing capabilities. Flexibility is another noteworthy aspect of Edge Impulse. The platform supports a wide array of devices, including microcontrollers, single-board computers, and mobile devices. This versatility empowers developers to build AI-powered applications for diverse edge devices. Latency is dependent on the specific deployment device and the complexity of the machine learning model being employed.

Edge Impulse provides comprehensive data processing capabilities, encompassing data collection, preprocessing, and feature extraction. It also supports various sensor types, ensuring compatibility with a range of data sources. The platform offers multiple pricing plans for professional product development and a free option for developers working on individual projects.

The ability to shrink a model's memory consumption while retaining its accuracy and the resulting efficiency achieved are core considerations in Edge Impulse's design, enabling inference to be performed on low-power devices. The platform seamlessly integrates with a wide range of edge devices, including microcontrollers, single-board computers, and mobile devices. The size of the deployment depends on the chosen device and the dimensions of the machine-learning model.

Integration with popular development tools and frameworks, such as Arduino, TensorFlow, and PyTorch, is a seamless process with Edge Impulse. This ensures a smooth incorporation of machine-learning capabilities into existing workflows, streamlining development processes for users and organizations.

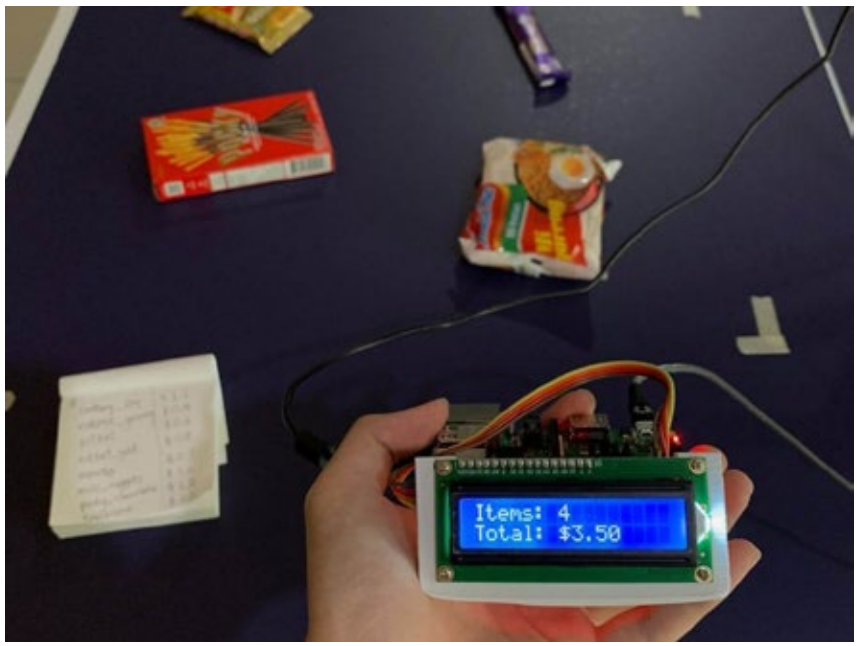
Caffe2

Caffe2 is an open-source deep learning framework that plays a significant role in developing Edge AI applications. Its prominent advantage lies in its high performance, enabling rapid and accurate processing of large datasets. Caffe2 is meticulously optimized for both CPUs and GPUs, and its compatibility with distributed computing environments ensures effortless scalability. This flexibility makes it adaptable to various computing scenarios.

Caffe2 stands out for its low inference latency, making it particularly suitable for real-time edge AI applications. Benchmarked tests have demonstrated its impressive performance on both CPU and GPU platforms.

For instance, Facebook AI Research conducted a benchmark test that showcased mean inference times of 14.4 ms on a mobile CPU and 4.8 ms on a GPU.

Efficient processing of large datasets is made possible by Caffe2's support for distributed training and automatic parallelization of operations. Its versatility also extends to data formats, with comprehensive support for images, audio, and video. Consequently, Caffe2 is an excellent choice for training and deploying large-scale machine-learning models.



The Edge Impulse FOMO (Fast Objects, More Objects) enables microcontrollers to perform object detection while running faster and consuming far less RAM. Image credit: Hackster.io.

Being an open-source framework, Caffe2 is freely available, making it an affordable option for developers and organizations seeking to build edge AI applications without incurring substantial software costs. However, it's important to note that Caffe2's power consumption tends to be higher due to its extensive use of GPUs.

Caffe2 boasts a compact code footprint, making it easy to deploy on edge devices with limited storage capacity. The library size of Caffe2 is less than 3 MB, making it an ideal choice for resource-constrained devices. Furthermore, Caffe2 is optimized to minimize heat generation during inference, making it suitable for edge devices operating in demanding environments. By efficiently dissipating heat, Caffe2 maintains peak performance even in high-temperature conditions.

Lastly, Caffe2 offers exceptional flexibility, allowing seamless integration with a wide range of hardware platforms, programming languages, and software frameworks. This versatility ensures compatibility and ease of use across diverse development environments.

MXNet

MXNet has gained significant popularity in machine learning due to its exceptional flexibility, scalability, and high-performance capabilities. One of MXNet's notable advantages is its extensive support for multiple programming languages, including Python, C++, and Julia. This feature empowers developers to work with their language of choice and facilitates seamless integration of MXNet into existing software systems.

Efficient data processing is another compelling aspect of MXNet. The framework's data processing engine is meticulously designed to handle large datasets efficiently. MXNet also supports data and model parallelism, enabling developers to scale their models across multiple GPUs or machines. Additionally, MXNet provides a rich collection of pre-built models and modules that accelerate the development of Edge AI applications.

MXNet boasts low latency and high-performance capabilities. Its advanced memory management system and deep learning optimization algorithms enable the swift execution of complex models. MXNet's focus on performance ensures that inference and training tasks are carried out with remarkable efficiency.

From a cost perspective, MXNet is an open-source framework, making it freely available for use and distribution. This characteristic makes MXNet an affordable option for organizations seeking to leverage machine learning capabilities without incurring substantial expenses.

Regarding power consumption, MXNet's optimized algorithms and memory management system contribute to its energy efficiency. This attribute is crucial for Edge AI applications that rely on battery or renewable energy sources.

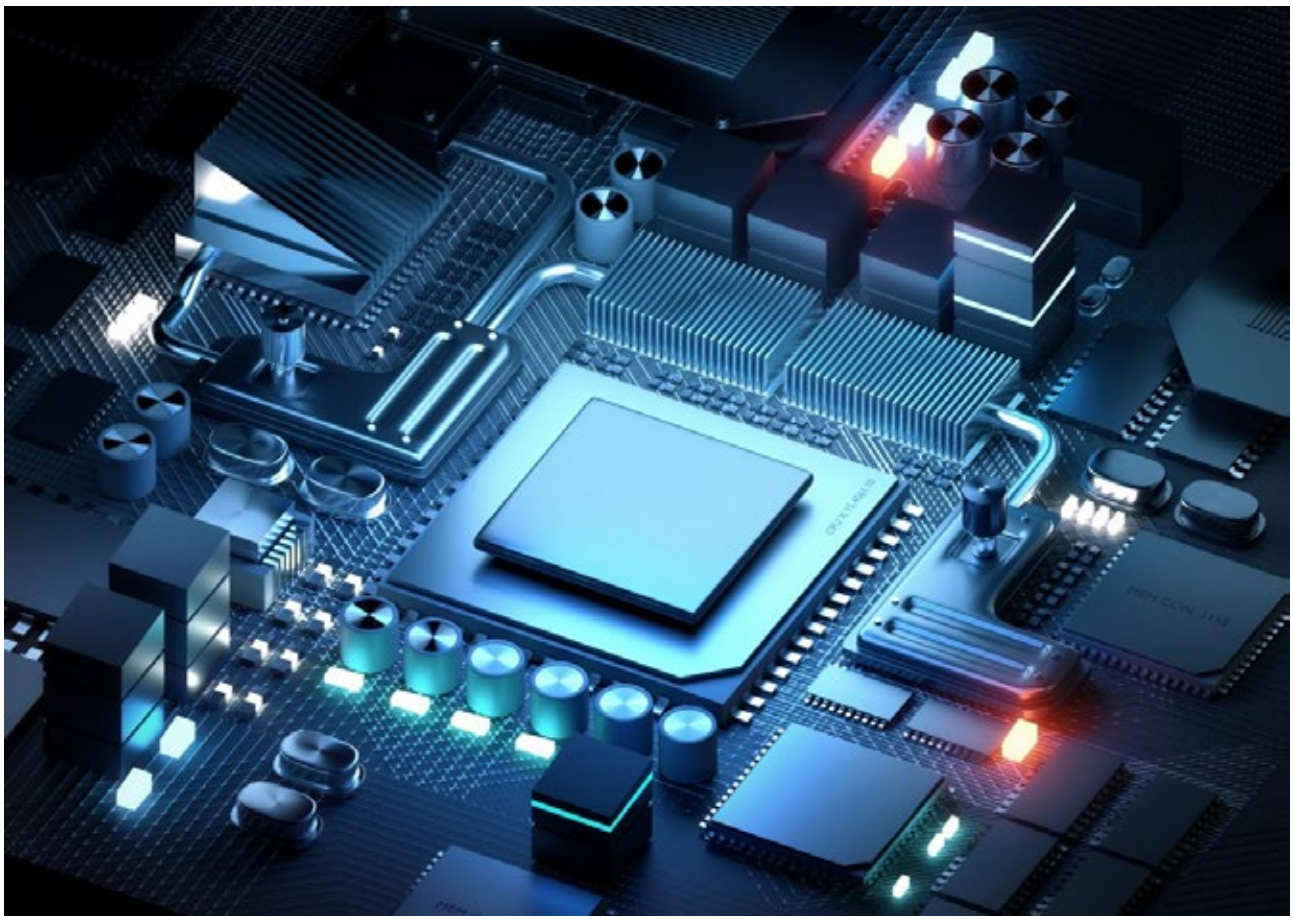
MXNet also excels in terms of size and weight. With its lightweight nature, MXNet can run efficiently on low-power devices like Raspberry Pi. This makes it an attractive choice for Edge AI applications that demand compact and portable systems.

Furthermore, MXNet's functional integration capabilities are noteworthy. It can seamlessly integrate with other deep learning frameworks like TensorFlow and PyTorch, allowing developers to leverage their preferred framework while still benefiting from MXNet's unique features and capabilities.

Edge AI platforms play a crucial role in enabling the development and deployment of machine learning models on edge devices. These are just a few examples of the diverse platforms available to developers. By carefully considering factors such as project requirements, performance capabilities, level of expertise, support, ecosystem compatibility, and cost, developers can select the most suitable platform for their specific Edge AI projects. It is important to remember that Edge AI is constantly evolving, and new platforms and technologies are continuously emerging. Staying updated with the latest advancements and conducting thorough evaluations will enable developers to harness the full potential of Edge AI platforms and drive innovation in this exciting field. With the right platform at their disposal, developers can unlock new possibilities and create intelligent edge applications that revolutionize industries and enhance our daily lives.

Chapter IV: Hardware & Software Selection

When selecting and implementing Edge AI systems, it's important to consider the limitations of edge devices and their compatibility with other hardware and software. Edge devices like sensors and smartphones often have limited processing power, storage, and battery life. These limitations can negatively affect the performance of Edge AI systems, as they may require significant resources to perform complex tasks, such as image recognition and natural language processing. To overcome these limitations, AI models must be optimized to work with limited resources, and specialized hardware such as field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), and GPUs can accelerate computation at the edge. Edge AI systems are designed to operate for extended periods, so the hardware selected should be capable of supporting future upgrades. It should be designed to accommodate new technologies and features that may be required in the future. Additionally, compatibility is a critical consideration when selecting software for Edge AI systems. Because Edge AI systems often require integration with existing infrastructure and hardware, it's important to select software that is compatible with a wide range of devices and platforms. This includes both hardware compatibility, such as compatibility with different types of sensors and processors, and software compatibility, such as compatibility with different operating systems and programming languages.



A field-programmable gate array (FPGA). Image credit: Microchip.

Hardware Considerations

Ensuring that the chosen hardware meets the requirements of Edge AI systems involves considering various factors. One key consideration is processing power, as Edge AI applications require hardware capable of processing data quickly and accurately. The hardware's computing power and clock speed play crucial roles in achieving real-time processing, which is essential in Edge AI systems. Another critical consideration is the amount and type of memory available in the hardware. Edge AI applications necessitate sufficient memory to store and process large volumes of data, with a focus on fast and real-time processing capabilities. There is a race to have the highest number of TOPS (tera-operations per second) in the hardware industry, but it's not a one-to-one comparison.

"The memory system handling, DMA configuration, software compression of weights, intelligence, and decompression all matter significantly. Focusing only on the box numbers is similar to focusing solely on the megapixel or gigapixel of a phone's camera without considering other factors. This tunnel vision is an error for the industry to make."

Rahul Venkatram, Senior Product Manager, Machine Learning, Arm

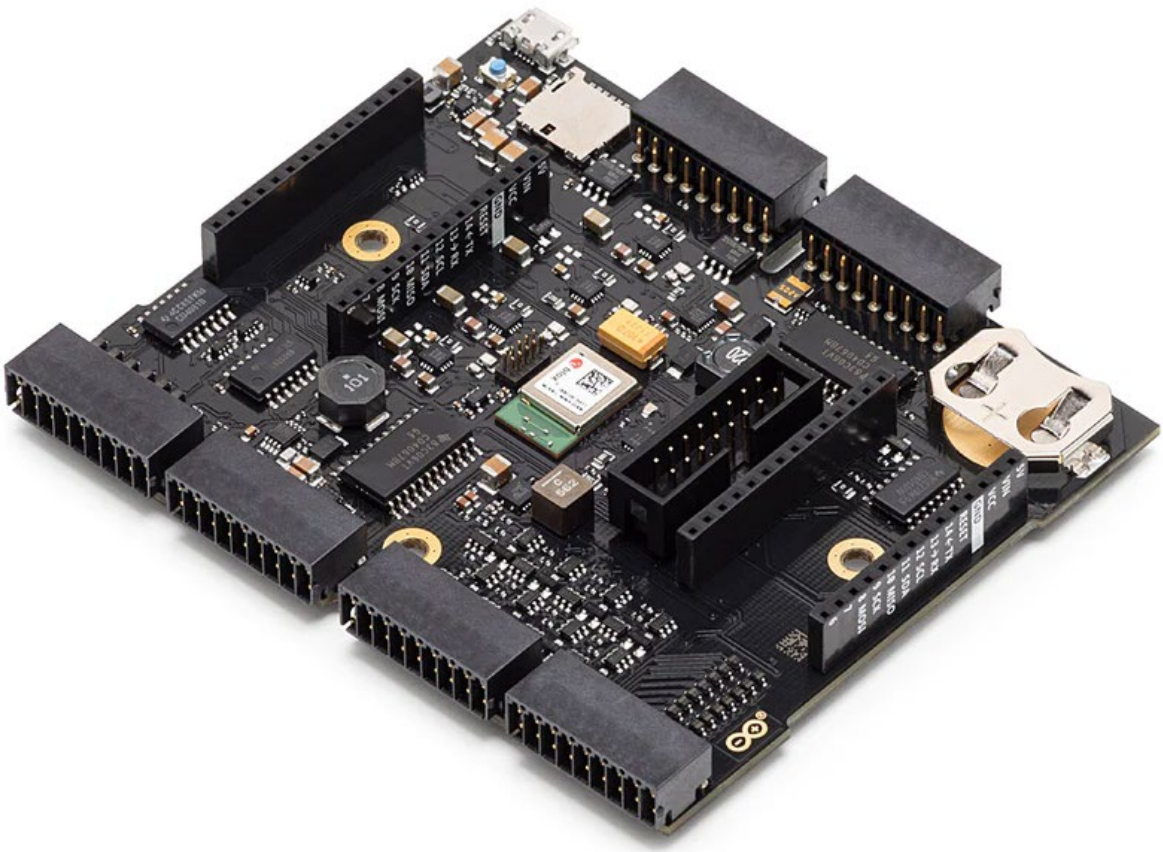
Power consumption is another vital aspect, particularly for devices with limited power resources like IoT devices. Hardware optimization for low power consumption becomes crucial to ensure prolonged device operation without frequent battery replacements or recharging. Techniques such as utilizing low-power chips, hardware accelerators, and intelligent power management systems contribute to achieving low power consumption in Edge AI systems. In a conversation with Henrik Flodell, Director of Product Marketing at Alif Semiconductor, he emphasized that "reducing power consumption is a key parameter users need to consider when they pick suitable devices for edge systems that will use ML."

The Autonomous Intelligent Power Management (aiPM™) technology from Alif is a vital solution for power consumption. It helps with selectively powering on the sections of an SoC that are needed when they are needed, and it powers them off when they are not, based on instantaneous processing demand and use case. It makes use of dual-processing regions within the MCU to speed up machine learning. The always-on, high-efficiency region continuously senses the environment, while the high-performance region wakes up when needed so it can execute heavy workloads and then return to sleep. This approach allows powerful devices to behave like low-power MCUs when necessary, which enables longer performance of smart IoT devices on smaller batteries.

"Systems that have to rely on CPU-bound processing tend to be power hungry, as the inference operations would take longer to finish, and the controller would run at 100% for much longer periods of time. Well-designed modern platforms offload the ML processing on accelerators designed for such workloads and can, therefore, finish faster and go to sleep sooner."

Henrik Flodell, Direct of Product Marketing, Alif Semiconductor

Seamless connectivity is also essential for Edge AI systems, requiring hardware that enables smooth data exchange with other devices and cloud-based platforms. The hardware should offer various connectivity options, including Wi-Fi, Bluetooth, and cellular networks, facilitating efficient communication and data processing with other devices. This connectivity ensures that data is processed and analyzed quickly and efficiently, contributing to the overall performance of the Edge AI system.



The Edge Control from Arduino is a remote monitoring and control solution optimized for outdoor environments that enables users to collect real-time data from smart sensors and leverage AI at the edge. Image credit: Arduino.

Software Considerations

When it comes to Edge AI systems, selecting software with specific characteristics is crucial for optimal operation. Analyzing various factors ensures that the chosen software meets the requirements of the Edge AI system. Compatibility stands as a significant consideration in software selection. The software should run efficiently on the hardware, enabling real-time data processing. Moreover, compatibility with other software components used in the system, such as operating systems, libraries, and frameworks, ensures seamless integration and functionality.

Scalability is another vital aspect to consider in Edge AI system design. Given that Edge AI systems often handle large amounts of data, the software must be capable of handling the real-time processing and analysis demands associated with such data. Scalable software guarantees the system can accommodate increasing data volumes, processing requirements, and user requests without compromising performance.

The accuracy of the software employed in Edge AI systems holds great importance. These systems heavily rely on accurate data analysis and processing to provide meaningful insights and support decision-making. Therefore, the software must exhibit high accuracy and reliability in analyzing and processing data.

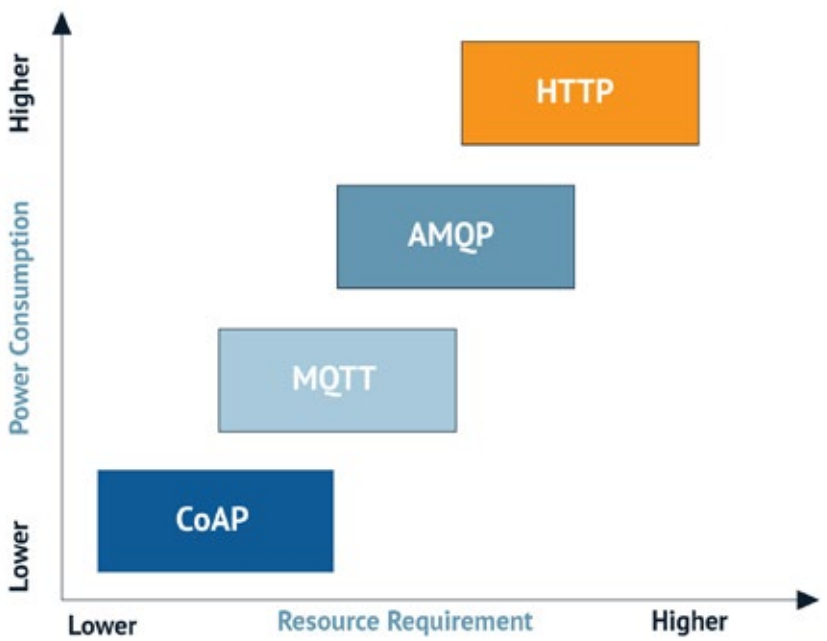
Interpretability, or the software's ability to explain its results comprehensibly, plays a crucial role in Edge AI system design. Interpretability allows users to understand the system's decision-making process and provides insights into data analysis. This aspect becomes particularly significant in applications where the decisions made by the Edge AI system have substantial implications, such as healthcare and finance. The software used in Edge AI systems should prioritize interpretability, presenting results in a clear and understandable manner.

Integration Considerations

Integrating Edge AI with other systems is essential for seamless and efficient operations. As Edge AI systems generate vast amounts of data that require processing, analysis, and sharing with other systems, selecting communication protocols becomes imperative. Different protocols are used by Edge AI systems to communicate with other systems, and it is crucial to choose a protocol that is compatible with the integrated systems.

Several commonly used protocols in Edge AI system integration include MQTT, CoAP, and HTTP. MQTT is a lightweight protocol widely utilized in IoT systems due to its low overhead and power consumption. CoAP, on the other hand, is designed specifically for resource-constrained networks, making it suitable for Edge AI systems. HTTP, a widely used protocol in web-based applications, is also suitable for Edge AI systems that communicate with web-based applications. The choice of communication protocol should align with the requirements of the Edge AI system and the systems being integrated.

In addition to communication protocols, data formats are another critical consideration when integrating Edge AI with other systems. Edge AI systems generate data in various formats, and it is vital to ensure compatibility between the data generated by the Edge AI system and the systems it interacts with. Data formats must be aligned and properly transformed or translated to facilitate seamless data exchange and interoperability between the systems.



Comparison of protocols in terms of power consumption and resource requirement. Image credit: Naik, N., IEEE International Systems Engineering Symposium (ISSE), 2017.

Global technology provider, Arm, has approached integration from a holistic thought process about where and how to deploy processing. David Maidment, Senior Director, Secure Device Ecosystem at Arm, explained their approach as follows: "We've landscaped our approach in two directions. The first is scaling down traditional cloud-native workloads, such as microservers. The second direction is taking IoT and embedded devices and making them connected, updateable, and secure. We're seeing these two trends converge, with higher-end, higher-power devices coming down in cost and power, and lower-end, lower-power devices being pushed up in performance capabilities."

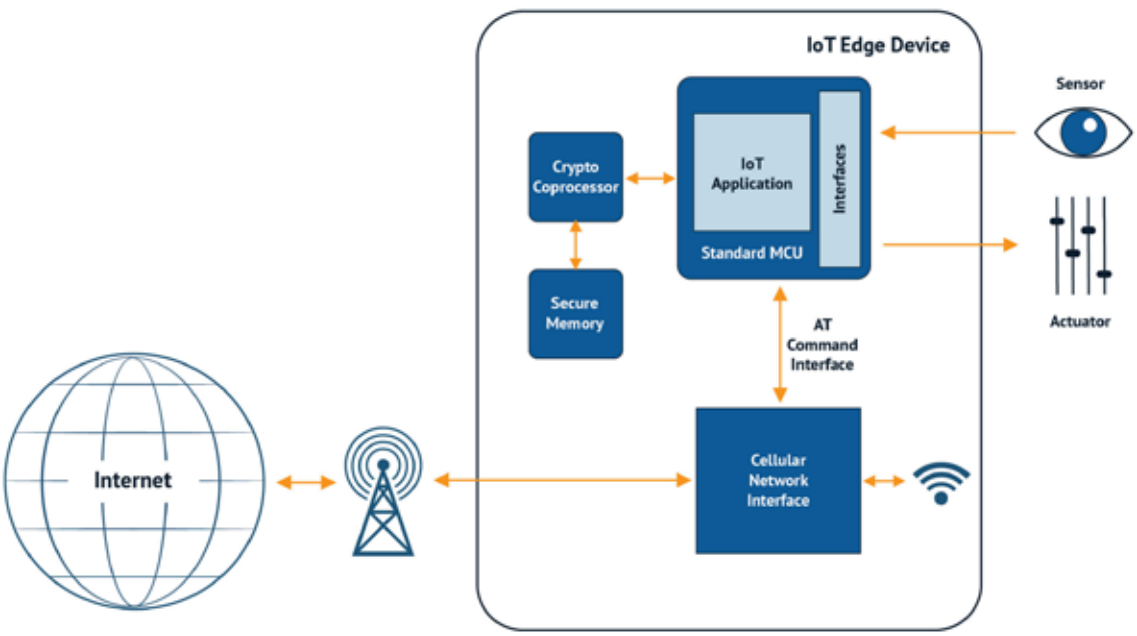
This results in two distinct software ecosystems. The first is the traditional cloud software ecosystem, with portable workloads that can run on servers with offload and acceleration. The second is a more traditional Linux box that was initially designed for a single function, like a router or gateway but is now being developed as a connected general-purpose box with capabilities that can change during its lifetime through new applications and AI models.

Security Considerations

Ensuring security is a critical factor when integrating Edge AI with other systems. Since Edge AI systems generate vast amounts of data, including sensitive information, protecting it from unauthorized access becomes imperative. Secure communication protocols such as SSL/TLS and SSH should be employed to establish secure channels for data transmission between the Edge AI system and other systems. These protocols encrypt the data, preventing eavesdropping and tampering during transit.

To reinforce security, access controls and authentication mechanisms must be implemented. This ensures that only authorized users can access the data generated by the Edge AI system, safeguarding it from unauthorized individuals. Secure software development practices, such as regular code reviews and vulnerability assessments, should also be followed to identify and mitigate potential security risks in the integrated systems.

Encryption is pivotal in securing Edge AI systems and should not be overlooked. By converting sensitive data into an encoded form, encryption ensures that even if the data is compromised, it remains unintelligible to unauthorized entities. Given that Edge AI systems often process sensitive data like personal health information, financial records, and valuable intellectual property, encrypting this data is crucial in preventing unauthorized access, theft, or manipulation.



Key usage in a cryptographic system of an IoT device and attack paths.

Chapter V: TinyML

Introducing TinyML

Edge AI systems can be deployed in various configurations and across many different types of devices and data sources. For instance, a typical Edge AI configuration involves deploying machine learning models in local clusters of computers, Internet of Things (IoT) gateways, and fog nodes. Such configurations deliver most of the benefits of the Edge AI paradigm, as they improve the latency, performance, energy efficiency, and security of AI applications.

However, there are also Edge AI applications that deploy machine learning models within embedded systems, such as IoT devices. Such configurations fall into the realm of embedded machine learning.

Embedded systems are typically part of another device (e.g., an automobile GPS unit or a smart home device) and perform specialized functions within those products. Embedded machine learning can therefore enhance the intelligence of such functions.

In recent years, there has been a growing interest in TinyML, a novel class of embedded machine learning systems that deploy AI models on microchips and microcontrollers (MCUs) to enable extremely low-power on-device data analytics. MCUs are small single-board computers built around one or more processors with limited memory and computing capacity. As Martin Croome, VP of Marketing at GreenWaves Technologies, explains, "TinyML targets devices and applications that are extremely energy-constrained. Generally, the device is expected to run on a milliwatt or lower power range." Devices of this type may be powered by batteries or energy harvesting and are expected to have long battery lives. The specific power constraints drive a broad range of technologies in algorithms, software tools, specific hardware, and sensor technologies. Very often, the requirement for TinyML is combined with a need for ultra-low-power digital signal processing. Examples can range from sensors that are expected to last for several years on batteries to medical devices like pill cameras or consumer products like earbuds, which have tiny batteries that need to last for hours.

With Tiny, it is possible to execute sophisticated machine learning models such as deep neural networks on them. This enables a variety of innovative and always-on use cases, such as on-device image recognition, object tracking, and real-time event detection for security applications. In IoT sensor applications, TinyML is used to interpret rich data sources such as images, sounds, or vibrations. Doing inference on-device reduces the large amounts of data down to compressed metadata such as the presence of and location of a specific object in a picture or a specific sound type. This alleviates the need to transmit the original raw sensor data off of the device, which reduces energy consumption (allowing battery operation) and improves security and privacy. In some cases, the latency required between 'seeing' a particular event and taking action precludes transmitting and processing data off of the device.

TinyML is also used in consumer electronics devices such as earbuds. Here more sophisticated deep learning models can be used to analyze incoming sounds, such as the user's voice, and generate filters that can remove unwanted and disturbing noise during calls. This type of neural network belongs to a class that is capable of separating different signals from a mix, such as removing the vocals or lead instruments from a song or separating different speakers.

TinyML techniques are also integrated into products with larger power sources to reduce energy consumption when in standby or idle modes.

TinyML Advantages and Challenges

The power consumption of TinyML deployments is in the order of milli-Watts (mW). This is usually the factor differentiating them from other instances of embedded machine learning. Hence, TinyML deployments drive the benefits of Edge AI to the extreme. Specifically, they tend to outperform other Edge AI deployments in the following aspects:

- **Latency:** TinyML systems do not transfer data to any server for inference, as machine learning functions are executed on the device itself. Thus, they save data transfer time, which optimizes latency as much as possible. Also, this makes TinyML very appropriate for certain types of real-time applications that require immediate feedback.

- **Power savings:** MCUs are extremely low-power devices that can operate for long periods of time without an energy source. This reduces the power needed for executing TinyML operations. Furthermore, these operations are remarkably power-efficient, given the lack of data transfers and I/O operations. These factors make TinyML deployments much more energy-efficient than other forms of Edge AI.

- **Bandwidth savings:** Unlike other Edge AI configurations, TinyML deployments do not rely on internet connectivity for inference. Data are captured and processed on the device rather than transferred to some server. Thus, TinyML systems end up being extremely bandwidth-efficient, as well.

- **Stronger Privacy and Data Protection:** In the scope of a TinyML deployment, there are no data on fog nodes and other servers. This makes it impossible for malicious actors to access sensitive information by hacking edge servers or networking devices. Overall, the execution of machine learning models on an embedded device leads to stronger data privacy guarantees.

- **Increased Efficiency and Flexibility:** MCUs and other embedded devices are smaller in size than servers and personal computers. This means they require much less space and power while providing increased deployment flexibility. For instance, TinyML systems offer one of the best ways to deploy machine learning models in places where there is not enough space for bulky equipment.

Nonetheless, a few challenges and limitations still take place when developing, deploying, and running machine learning on embedded devices and microcontrollers. One of the main challenges relates to data collection processes. Given the limited amount of power available, it is quite difficult to collect large amounts of data for TinyML training tasks. This is also why there is limited availability of datasets for embedded machine learning and TinyML tasks.

Another challenge is the lack of computing capacity and memory for processing large amounts of information. This is generally a setback to implementing certain tasks (e.g., video scene analysis) within MCUs and other embedded systems.

TinyML deployments are also associated with a skill-gap challenge. Developing and deploying TinyML applications require multi-disciplinary teams that combine embedded systems and data science expertise. This is a challenging combination, considering the proclaimed gap in data science and embedded systems development skills.

Tools and Techniques for TinyML Development

The key algorithmic technologies for TinyML cover techniques to run large machine-learning models efficiently. These techniques cover model compression (reducing the size of the model) and model optimization (arranging the computation of the model in ways that reduce power consumption). Dedicated hardware is used to minimize computational energy use and costly movement of data. The interaction between sensor data collection, preprocessing, and inference in different applications and the speed of change in techniques means that there is a large need for hardware that preserves flexibility in the approach used to implement TinyML applications.

State-of-the-art TinyML processors and their toolchains are able to process deep neural networks such as object detectors that previously consumed 100s of milliwatts or even watts at power levels below a milliwatt.

The development and deployment of TinyML applications are based on machine learning, data science, and embedded systems programming tools. In principle, popular data science tools (e.g., Jupyter Notebooks, Python libraries, and tools) are used to train a machine learning model. Accordingly, the model is shrunk in size to fit the embedded device. In this direction, many popular machine learning tools

offer options and features for embedded development.

As a prominent example, a special edition of the popular TensorFlow suite of ML tools is designed for inference on devices with limited computing capacity, such as phones, tablets, and other embedded devices. This edition is called TensorFlow Lite, which is Google's lightweight, low-power version of TensorFlow.

Another special version of TensorFlow Lite for microcontrollers, namely TensorFlow Lite Micro, enables the deployment of models on microcontrollers and other devices with only a few kilobytes of memory. For instance, the core runtime of TensorFlow Lite Micro can fit in just 16 KB on an Arm Cortex M3.

Furthermore, embedded systems tools like the Arduino IDE can be used to transfer and execute small-sized models on the embedded device.

The training and development of TinyML pipelines are generally subject to the same model performance criteria as other ML pipelines. Nevertheless, TinyML developers must also consider additional factors prior to deploying a model to production.

For instance, they must consider whether the model's size fits the available memory and how well it performs in terms of execution speed. Likewise, they may also have to assess its impact on the device's battery life.

TinyML in Action: How GreenWaves Enable Next-Generation Products

Combining homogeneous processing units with integrated hardware acceleration blocks is designed to achieve a perfect balance between ultra-low power consumption, latency, flexibility, and ease of programming.

GAP9 from GreenWaves Technologies is a combination of a robust low-power microcontroller, a programmable compute cluster with a hardware neural network accelerator, and a sample-by-sample audio filtering unit.

The compute cluster is perfectly adapted to handling combinations of neural network and digital signal processing tasks delivering programmable compute power at extreme energy efficiency. Its architecture employs adjustable dynamic frequency and voltage domains and automatic clock gating to tune the available compute resources and energy consumed to the exact requirements at a particular point in time.

GAP9's unique Smart Filtering Unit is perfectly adapted to ultra-low latency (1uS) PDM to PDM filtering tasks but so flexible that it can simultaneously be used as a block filtering coprocessor for tasks executing on the cores. The SFU is linked to GAP9's 3 Serial Audio Interfaces, capable of handling up to 48 incoming or outgoing audio signals.

Its SDK allows a simple path from NN development packages, such as TensorFlow and PyTorch to C code running on GAP9.

GAP9's hierarchical and demand-driven architecture is focused on bringing signal processing with embedded artificial intelligence into the next generation of hearable products and applications for battery-powered smart sensors.

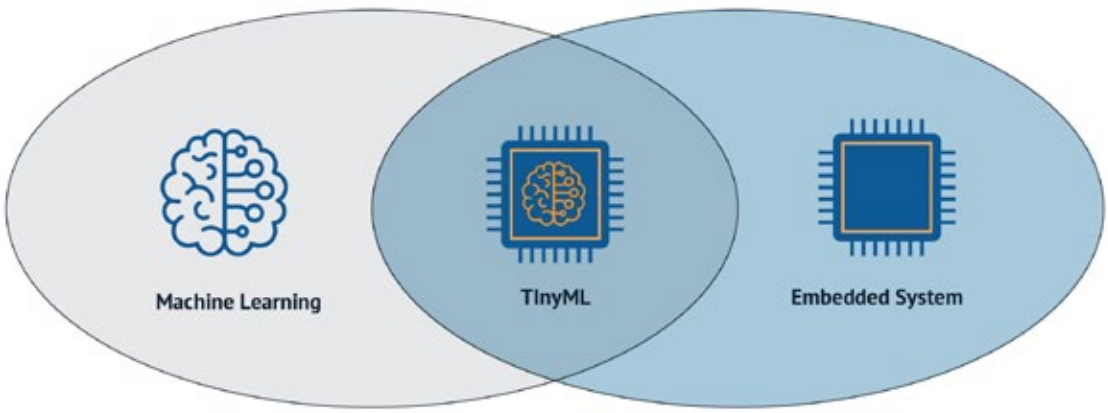
GAP9 processor, an advanced AI processor combining embedded AI and signal processing in wireless earbuds. Image credit: GreenWaves.

The New Kid on the Block

The rising popularity of TinyML has given rise to more sophisticated application development and deployment tools. For instance, many hardware vendors and OEMs (Original Equipment Manufacturers) offer visual AutoML (Automatic Machine Learning) tools, which ease the process of developing and deploying TinyML pipelines. In several cases, these tools offer options for monitoring the power consumption, performance, and speed of the TinyML deployment while facilitating their integration into other Edge AI systems in the cloud/edge computing continuum.

Shay Kamin Braun of Synaptics explains that TinyML has gained this popularity due to the potential advantages it brings to the industry and to various players in terms of application, design, and usability. As it is suitable for resource-constrained environments in both memory and processing power, TinyML allows us to design smaller products with lower costs and lower power consumption. "The lower costs of TinyML," clarifies Kamin Braun, "is enabled by utilizing lower-performance processors. And very importantly, lower power consumption leads to the ability to design devices that run on batteries and last a long time (i.e., multiple years, depending on the application)."

Nonetheless, TinyML is still in its infancy. As Kamin Braun told Wevolver, "We will see more of it in the near future, with applications in command detection (Wake word and phrase detection models), acoustic event detection (for security, smart homes, and industrial applications), and Vision AI (detecting people, objects, and movement) despite its minimal implementation today." TinyML's rather huge potential is the main reason for its popularity as compared to its actual implementation.



TinyML: a combination of machine learning and embedded systems.
Image credit: Leonardo Cavagnis. Adapted by Wevolver.

"We're just seeing the tip of the iceberg now. We expect exponential growth as more and more companies are creating small, high-performance models. Fast, low-cost deployment is slowly coming but not in mass yet. Maybe it willin a year or so."

Shay Kamin Braun, Director of Product Marketing, Synaptics

Chapter VI: Edge AI Algorithms

Artificial Intelligence (AI) deployment on edge devices has become increasingly popular due to its ability to perform tasks locally without relying on cloud services. However, one of the most critical considerations is the selection of a suitable algorithm that is appropriate for the problem it is intended to solve. The complexity, size, and accuracy of AI models can vary significantly, and choosing the best-performing algorithm may not be enough.

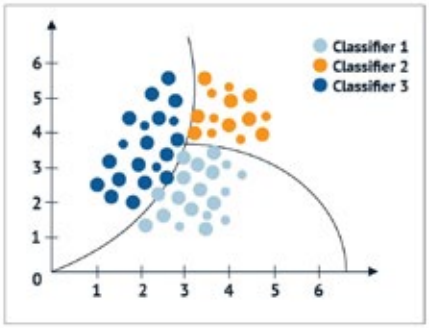
In many cases, edge devices have limited computing resources, so the selection of the algorithm should also take into account the computation power of the edge device and whether the chosen algorithm can run smoothly on the hardware. Balancing the algorithm's performance with the available resources is essential to ensure that the edge device can execute the AI model effectively. While various types of AI algorithms can be successfully utilized on edge devices, including regression and classification algorithms, recent advancements and practical applications have shown that other types of applicable algorithms, such as clustering algorithms and natural language processing (NLP) algorithms. Clustering algorithms enable edge devices to group data points into clusters based on their similarities, while regression algorithms can predict the relationship between different data points. NLP algorithms allow edge devices to understand and respond to natural language commands.

However, the most popular and suitable algorithms for deployment on edge devices are classification, detection, segmentation, and tracking algorithms. These four algorithm types offer practical solutions for various applications, from object recognition and tracking to quality control and predictive maintenance. Here, we will discuss these four algorithm types in detail and explore their practical applications on Edge AI systems.

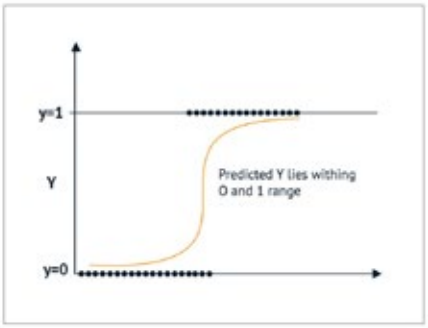
Classification Algorithms

Classification algorithms play a significant role in Edge AI technology as they enable edge devices to recognize and categorize different types of data. Recognizing and classifying data is essential for many Edge AI applications, such as object recognition, speech recognition, and predictive maintenance. Using classification algorithms, edge devices can process data locally, reducing latency and saving on network bandwidth. Several commonly used classification algorithms are applied in edge computing, including Support Vector Machines (SVMs), Decision Trees, Random Forests, and Convolutional Neural Networks (CNNs). We'll focus on these algorithms next.

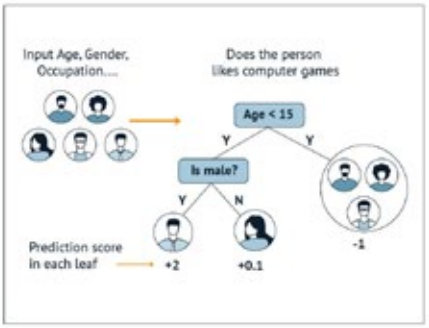
Machine Learning classification algorithms.
Adapted from Omdena by Wevolver.



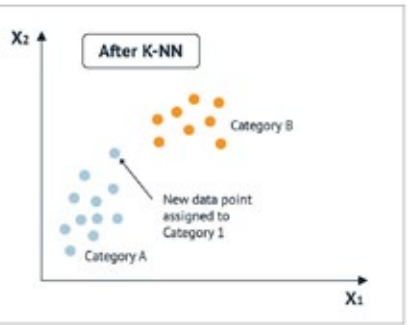
Naive bayes classifier



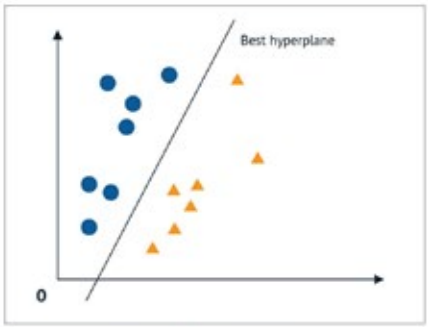
Logistic Regression



Decision Tree



K-Nearest Neighbors



Support Vector Machines

Support Vector Machines

SVMs are popular binary linear classifiers used in classification tasks such as image recognition. These algorithms work by identifying a hyperplane in high-dimensional space that can best separate the different classes of data. The distance between the hyperplane and the closest data points from each class is maximized to increase the classification accuracy. SVMs are particularly well-suited for smaller datasets and have a strong theoretical background for binary classification problems. They can perform well even with high-dimensional features, whereas other methods may face difficulties or limitations. However, SVMs can be sensitive to the choice of the involved kernel function, which maps the data into a higher-dimensional feature space. The choice of this function can significantly affect the model's accuracy and computation time, so precise selection and efficient optimization are necessary to achieve desired performance.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees to make predictions. Each tree is trained on a random subset of the data and produces its prediction, and the final prediction is determined by aggregating the predictions of all the trees in the forest. Random Forest is an excellent solution for handling large datasets and high-dimensional features because it can handle a large number of input features without overfitting the model to the training data. This is possible because the trees in the Random Forest are trained on different subsets of the data and different subsets of the features, reducing the risk of overfitting. Random Forests are particularly useful for classification problems with a large number of input features, such as image classification tasks.

Convolutional Neural Networks

CNNs are deep learning algorithms that excel at image recognition tasks by effectively extracting features from raw input data. They exploit a series of convolutional layers to learn filters applied in feature extraction. The pooling layers that follow these convolutional layers reduce the spatial dimensions of the feature maps and introduce invariance to small translations in the input. The scalability of CNNs is one of their strengths, as they can handle a wide range of image sizes and resolutions. In addition, CNNs can be trained on large datasets using stochastic gradient descent and backpropagation to learn highly complex and abstract features. To optimize CNNs for low-power applications, techniques like weight pruning, quantization, and model compression can be used. This makes CNNs an ideal solution for edge devices in energy-constrained environments.

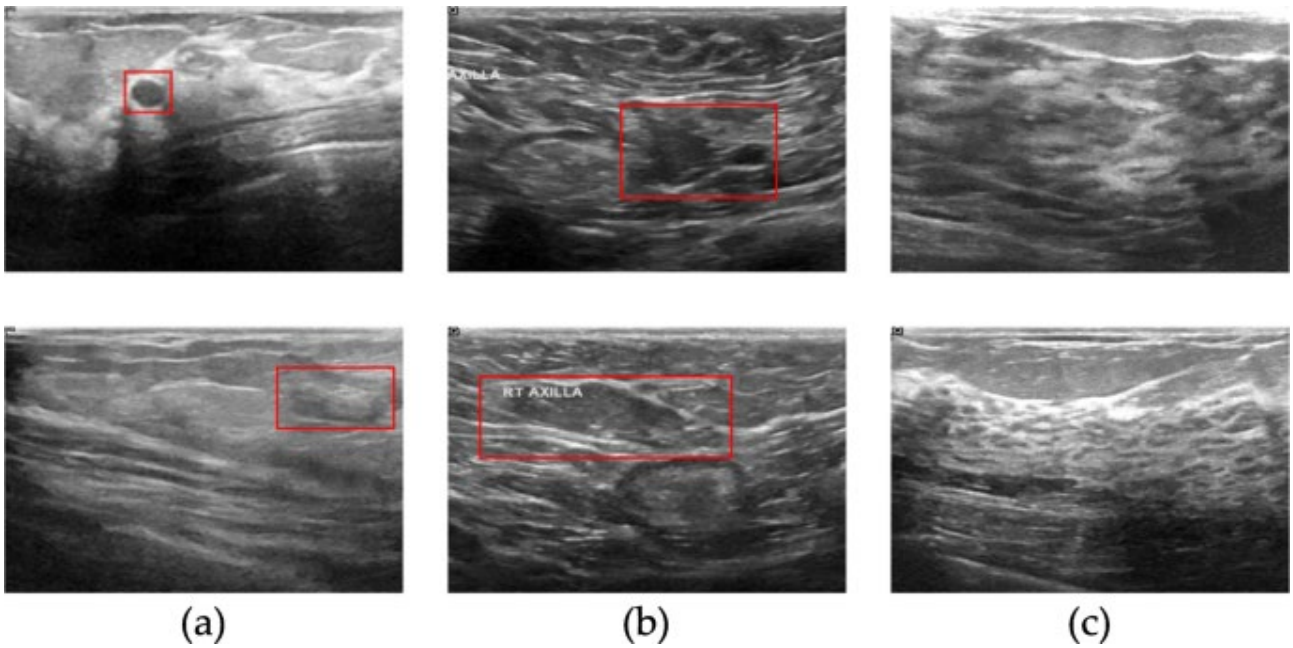
Detection Algorithms

Detection AI algorithms play a crucial role in Edge AI systems by facilitating real-time analysis and decision-making at the edge of the network. These algorithms leverage sophisticated machine learning techniques, including computer vision, natural language processing, and signal processing, to identify specific patterns or events in real time. By deploying detection AI algorithms, Edge AI systems can operate with low latency and high accuracy without relying on frequent communication with a cloud provider. This is particularly advantageous in applications where real-time analysis is vital, such as security and surveillance systems, industrial automation, and autonomous vehicles. Furthermore, deploying such algorithms can reduce the data transfer requirements and associated costs of Edge AI systems since only relevant data is transmitted to the cloud for storage. The most commonly used categories of detection algorithms include object detection, anomaly detection, event detection, and face detection algorithms.

Object-Detection Algorithms

Object-detection algorithms are essential tools in computer vision, enabling accurate object recognition within images or videos. These algorithms identify regions of interest in an image or video frame and classify them based on the objects they contain. The accuracy of these algorithms relies on the quality and size of the training dataset, as well as the complexity and performance of underlying AI models.

Deep learning-based object detection algorithms, such as YOLO and Faster R-CNN, have shown good performance in various applications, including robotics, autonomous vehicles, and surveillance systems, where they are employed for real-time object detection. Additionally, the algorithms are successfully used in the medical field to help identify anomalies in medical images. However, detecting small or occluded objects remains a challenge for most algorithms, requiring new approaches and techniques to improve their robustness.



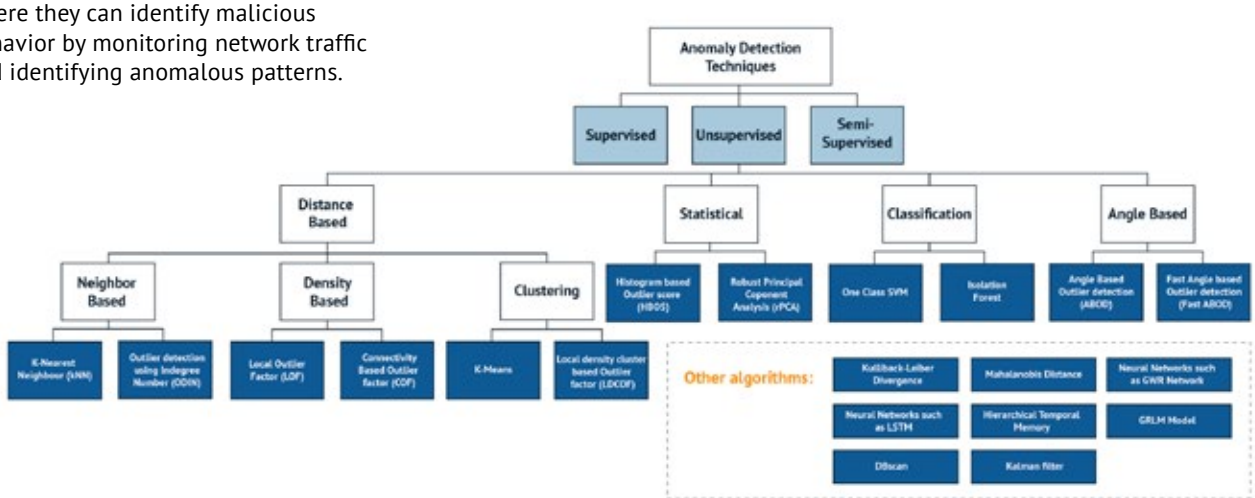
Anomaly detection model of mammography using YOLOv4-based histogram.
Image credit: Kim, C-M. et al., Personal and Ubiquitous Computing, 2023.

Anomaly-Detection Algorithms

Anomaly-detection algorithms are designed to identify and flag events that deviate significantly from a system's expected or normal behavior. These algorithms establish a pattern of normal behavior based on historical data and then monitor the system to compare new data against the established pattern.

There are various approaches to anomaly detection, including statistical methods, clustering algorithms, and deep learning. Deep learning-based anomaly detection algorithms, such as autoencoders and RNNs, have gained popularity because they can learn complex patterns and detect anomalies. However, one of the challenges of anomaly detection is the difficulty in defining what is expected or normal behavior for a given system, as this can vary depending on the application and the system itself.

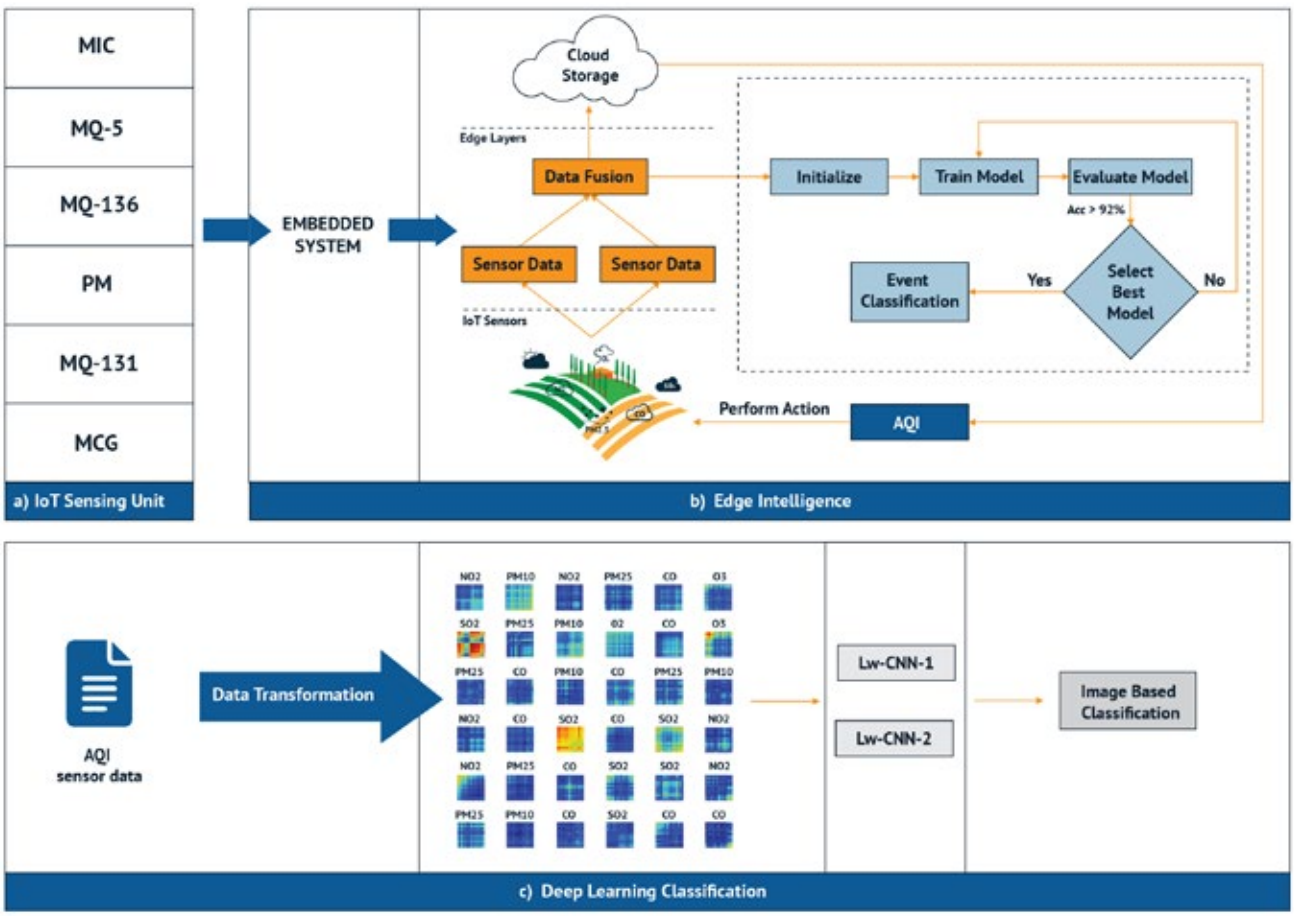
Anomaly detection algorithms are widely used in industrial applications for predictive maintenance, where they can identify equipment malfunctions before they occur, and in cybersecurity, where they can identify malicious behavior by monitoring network traffic and identifying anomalous patterns.



Anomaly detection techniques.

Event-Detection Algorithms

Event-detection algorithms can identify specific events or changes in data over time based on patterns or changes in the data. One commonly used algorithm for event detection is the background subtraction algorithm, which subtracts a static background image from each video frame and identifies pixels that differ from the background. Another popular algorithm for motion detection is optical flow, which tracks the movement of objects in successive video frames. For sound event detection, a widely used algorithm is Mel-frequency cepstral coefficients, which extract features from sound waves to identify specific sound patterns. These algorithms are highly valuable in various applications, such as security monitoring and industrial automation. For instance, motion detection algorithms can trigger an alert or begin video recording upon detecting suspicious motion. Similarly, sound event detection algorithms could identify events such as speech, music, or machine noise. In industrial automation, event detection algorithms can monitor manufacturing processes, detect when a machine has stopped working, and prevent potential damage situations.



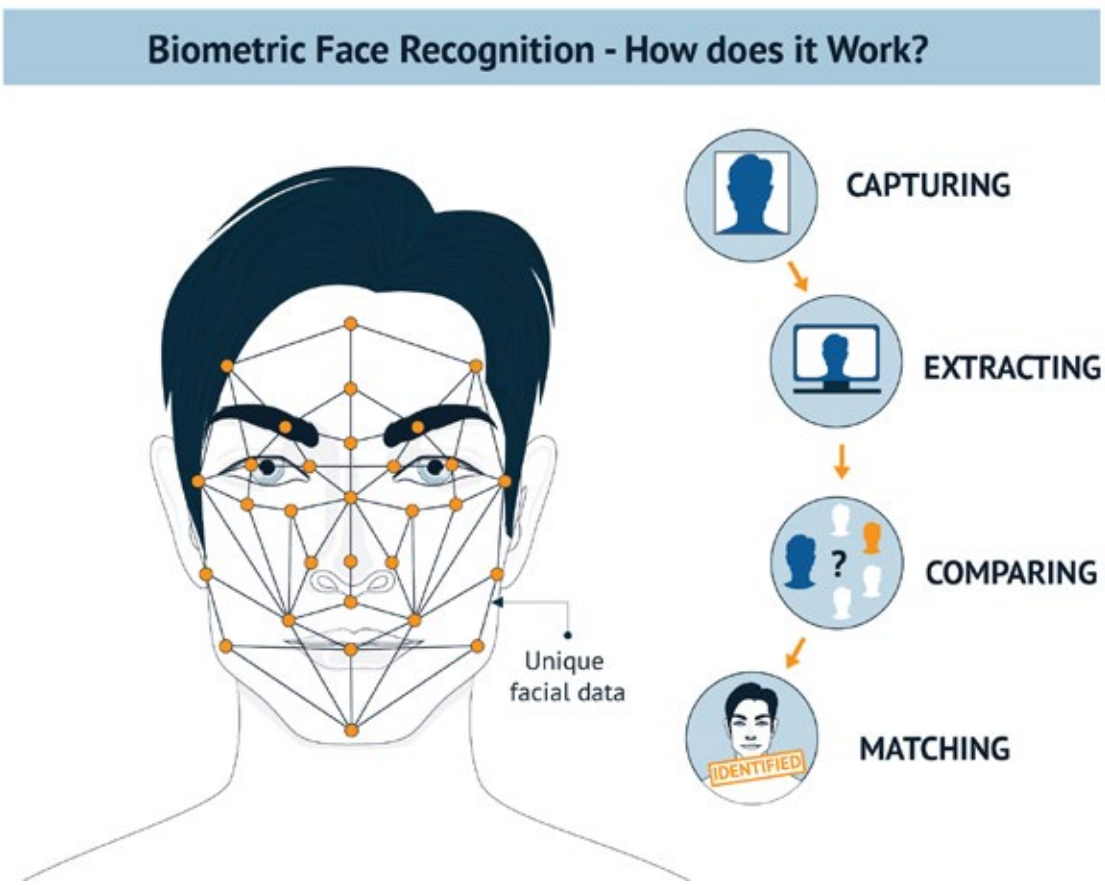
The architecture of event-based deep-learning framework for edge intelligence (EDL-EI). Image credit: Shah, S. K. et al., Sensors, 2021. Adapted by Wevolver.

Face-Recognition Algorithms

Face-recognition algorithms are biometric recognition systems that detect and identify human faces within a photo or video. These algorithms work by analyzing and comparing facial features, such as the distance between the eyes, the shape of the nose, and the contours of the face. The first step in face recognition is face detection, which involves identifying the presence of a face in an image or video frame.

Once a face is detected, the algorithm extracts its features and compares them to a database of known faces to determine the subject's identity. There are several approaches to face recognition, including geometric, photometric, and hybrid methods, where geometric methods analyze the shape and structure of the face, photometric methods analyze the patterns of light and shade in the face, and hybrid methods combine both geometric and photometric features to improve accuracy.

Face recognition algorithms are widely used in various applications, such as security and surveillance, law enforcement, and marketing. However, there has been an upsurge in scrutiny and regulation in some regions due to concerns about privacy and the potential misuse of facial recognition technology.



How biometric face recognition works.

Segmentation Algorithms

Segmentation algorithms enable real-time image segmentation, which divides an image into multiple segments corresponding to different objects or backgrounds. Image segmentation simplifies image representation, making it easier to analyze and interpret. Edge devices use segmentation algorithms to analyze images and video streams captured by cameras at the edge of the network. By segmenting images into distinct regions, these algorithms enable edge devices to detect and track specific objects in real-time. The popular segmentation algorithms include CNNs, Random Forest algorithms, and the K-Means Clustering approach.

CNNs extract features from images using convolutional filters and classify pixels in the image into regions. Random forest algorithms generate a segmentation map for the input image based on the output of decision trees. The K-Means Clustering algorithm first converts an image into a high-dimensional space represented by pixels. The objective is to minimize the sum of squared distances between each pixel and its corresponding cluster center by partitioning the image. This process iteratively updates the cluster centers until convergence.

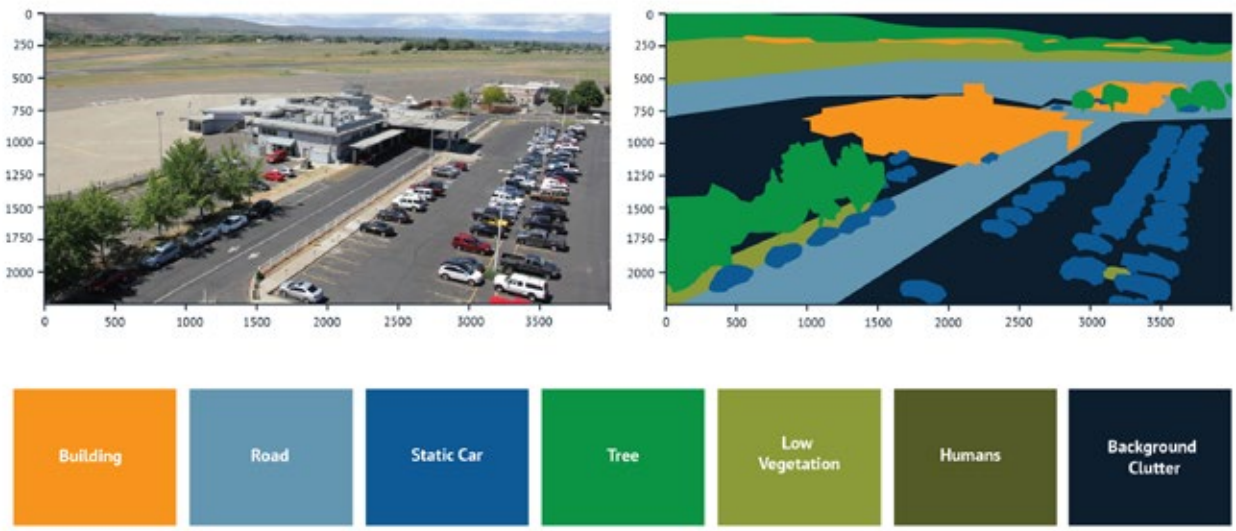


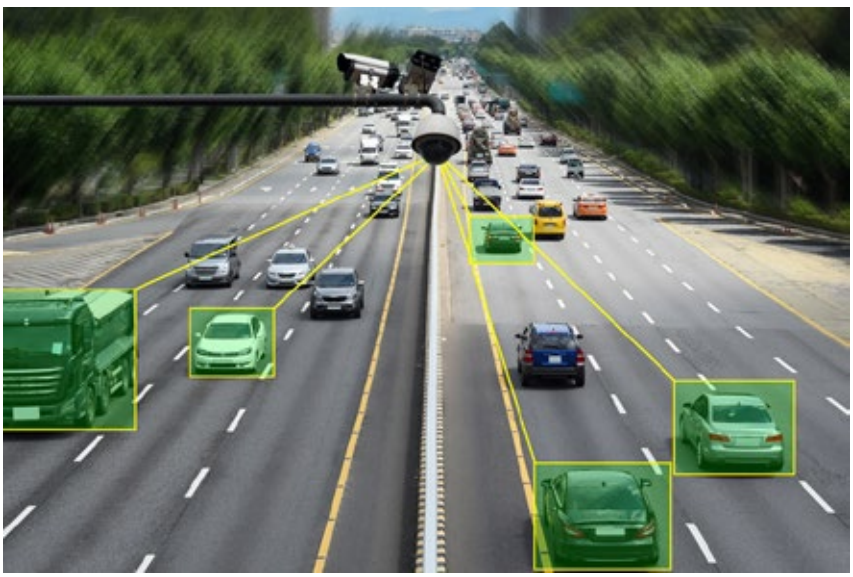
Image segmentation and classification using segmentation algorithms. Adapted by Wevolver.

Tracking AI Algorithms

The last group of AI algorithms we want to highlight are the tracking algorithms that enable real-time detection and tracking of specific objects. These algorithms are crucial for various applications, such as surveillance and autonomous vehicles. The primary objective of tracking algorithms is to locate and follow moving objects accurately, providing essential information for decision-making or triggering actions.

Kalman filters are a well-known solution for tracking objects in motion. These filters estimate the object's position, velocity, and acceleration and predict its future positions based on this information.

Particle filters are another popular solution that represents the object's position as a probability distribution and updates it over time based on new observations from images or videos. CNNs have also shown promising results in tracking tasks. They use convolutional filters to extract the features of the object of interest and track its movement. Furthermore, optical flow algorithms can identify movement patterns in an image and track the object using these patterns. These algorithms can track objects with smooth motion patterns, such as the motion of cars on a highway.



An AI-based video monitoring system that helps improve road infrastructure and traffic management.
Image credit: Advantech.

Temporal Event-based Neural Nets (TENNs)

Unlike standard CNN networks that only operate on spatial dimensions, TENNs are event-based networks that contain both temporal and spatial convolution layers. They may combine spatial and temporal features of the data at all levels, from shallow to deep layers. TENNs efficiently learn both spatial and temporal correlations from data in contrast with state-space models that mainly treat time-series data with no spatial components. Given the hierarchical and causal nature of TENNs, relationships between elements that are both distant in space and time may be constructed for efficient, continuous data processing. For applications like video, raw speech, and medical data, TENNs provide highly accurate processing with substantially smaller model size, and they can be trained just like CNNs.

Vision Transformers

An innovative advancement in Neural Networks for Computer Vision is the adoption of Vision Transformers (ViT) as a substitute for CNN backbones. Drawing inspiration from the remarkable performance of Transformer models in Natural Language Processing (NLP), researchers have begun applying similar principles to Computer Vision.

Prominent examples, including XCiT, PiT, DeiT, and SWIN-Transformers, highlight this trend. In this approach, images are treated as sequences of image patches, similar to NLP processing. Feature maps are represented as token vectors, with each token embedding a specific image patch. Vision Transformers are quickly gaining prominence in various Computer Vision applications. They outperform CNNs on extensive datasets due to their increased modeling capacity, reduced inductive biases, and wider global receptive fields.

Harmonizing Algorithms with Hardware

Edge devices often have limited computing power. So, choosing an algorithm that works well with the device's hardware is vital, balancing performance and resources.

Bram Verhoef, Head of Machine Learning at Axelera AI, explains how the combination of hardware and software design can be an effective solution to such a challenge. "This involves techniques like data compression, specifically data quantization and network pruning. By compressing the data, the device needs less storage and memory and can use simpler computation units, for example, those using fewer bits. This also improves energy efficiency."

"But data compression can cause the AI model to be less accurate," Verhoef adds. "To avoid this, developers need to carefully choose which parts of the AI network can be compressed and by how much without losing too much accuracy."

Additionally, the compression algorithm should be quick and compatible with many networks. Hardware developers need to work on using compressed data efficiently and communicate their findings to algorithm developers.

Chapter VII: Sensing Modalities

Sensing modalities enable capturing of environmental data for analysis using traditional and modern models and AI algorithms. By facilitating efficient data acquisition and processing, these cutting-edge technologies empower informed decision-making and significantly enhance productivity. The exploration of sensing modalities is of utmost importance, as it highlights their critical role in the comprehensive process of data collection, analysis, and, ultimately, intelligent decision-making.

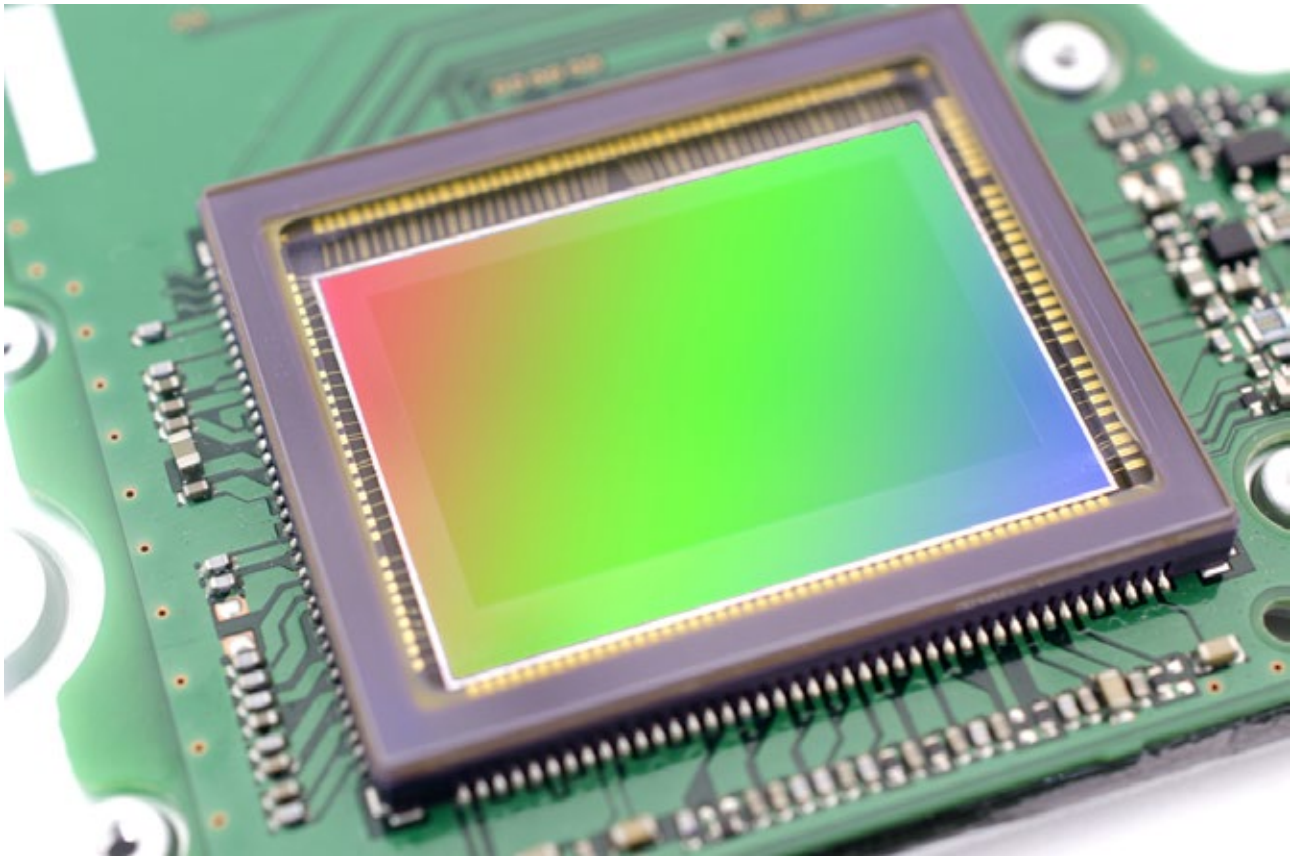
Sensing modalities encompass various sensors and data capture devices, such as cameras, microphones, temperature sensors, and motion sensors. Each modality specializes in capturing specific types of data, contributing to a holistic understanding of the device's operating environment. The significance of these modalities in Edge AI applications lies in their capacity to facilitate real-time data acquisition, ultimately enhancing device responsiveness and intelligence. By leveraging them, Edge AI systems can acquire valuable insights from the surrounding environment, enabling more informed decision-making and empowering devices to interact intelligently with their surroundings.

In terms of deployment, using sensing modalities at the edge enables devices to attain heightened situational awareness and make informed decisions, even in rapidly evolving environments. This attribute holds significant importance in applications that require quick response times, such as autonomous vehicles, industrial automation, and healthcare. Moreover, sensing modalities empower devices to capture data that would otherwise be difficult or impossible to collect. Environmental sensors, for example, provide insights into air quality and temperature fluctuations, while motion sensors track real-time object movement. By leveraging this diverse range of data, decision-making processes are enhanced, system performance is optimized, and energy efficiency is promoted. Here is a brief overview of the most common sensing modalities and their application value.

Vision-Based Sensing Modalities

Vision-based sensing modalities, employing sensors and cameras, play a vital role in Edge AI applications by capturing visual data and enabling sophisticated image and video processing techniques. These modalities encompass a range of technologies, including cameras and image sensors, which capture visual data from the environment.

By utilizing optical principles, these sensors convert incoming light into electrical signals, which are subsequently processed to generate images or videos. Cameras, the most common vision-based sensors, employ lenses to focus light onto an image sensor, such as a charge-coupled device (CCD) or complementary metal-oxide-semiconductor (CMOS) sensor.



A CMOS image sensor. Image credit: Association of Advanced Automation.

In the context of Edge AI, image and video processing techniques are crucial components of vision-based sensing. Image processing algorithms analyze individual images, while video processing techniques operate on sequences of images. These techniques encompass image enhancement, filtering, feature extraction, and compression. They are applied to individual frames or sequences of frames to detect patterns or anomalies in the data. Notably, deep learning techniques, such as convolutional neural networks, have demonstrated remarkable performance in image and video processing tasks.

With diverse applications in Edge AI, vision-based sensing modalities serve various purposes. Surveillance, for instance, relies on cameras to monitor specific areas, utilizing object detection algorithms to identify and track objects of interest, such as people or vehicles. Facial recognition, another vital application, employs cameras for face identification, serving security and access control purposes. Additionally, object tracking utilizes cameras to monitor object movement in specific areas, finding applications in industrial process monitoring and vehicle tracking in traffic. Vision-based sensing modalities also involve object recognition and detection algorithms for identifying objects in visual data. These algorithms utilize features extracted from the data to match objects with known models or templates. Object recognition algorithms are applied in facial recognition applications, while object detection algorithms find use in surveillance and object-tracking tasks.

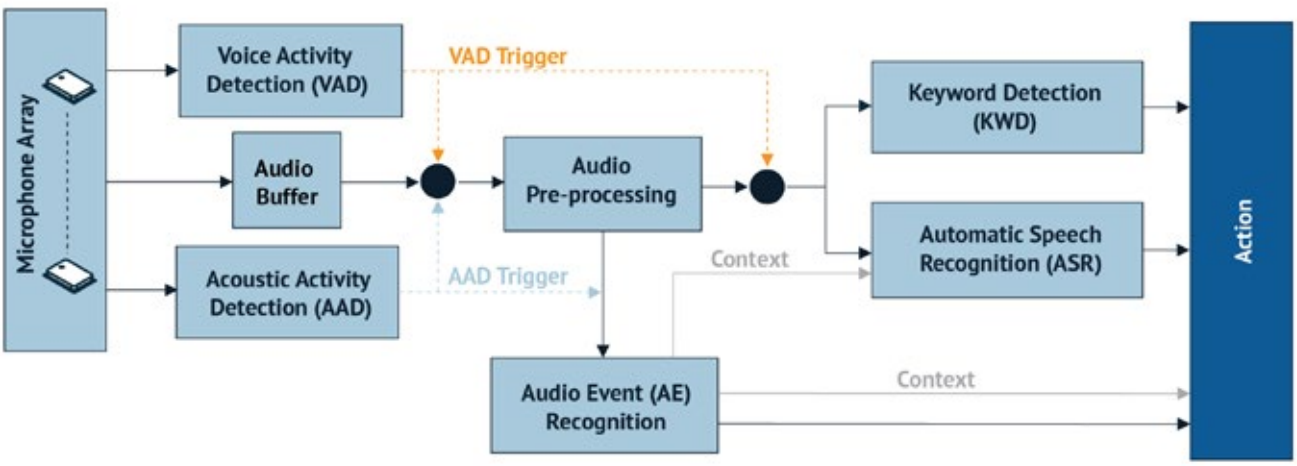
Audio-Based Sensing Modalities

Audio-based sensing modalities are essential for acquiring and analyzing audio data from the surrounding environment. The primary devices used for audio data acquisition are audio sensors and microphones, which come in various types, such as dynamic, condenser, and ribbon microphones, and can be customized to fit specific use cases. These components capture sound waves in the environment and convert them into electrical signals for further processing.

Audio sensors can vary in design and characteristics, including omnidirectional or directional sensitivity, frequency response, and signal-to-noise ratio. On the other hand, microphones encompass a wide range of types, such as condenser microphones, dynamic microphones, and MEMS (Microelectromechanical Systems) microphones, each offering distinct advantages for different applications.
Another significant application of audio-based sensing is acoustic monitoring. By strategically placing microphones in diverse environments, acoustic monitoring systems can detect and analyze sounds for multiple purposes. In industrial settings, acoustic monitoring plays a crucial role in identifying machine malfunctions or anomalous sounds, enabling predictive maintenance and improving operational efficiency. Similarly, in healthcare, acoustic monitoring systems assist in patient monitoring by detecting events such as snoring, coughing, or abnormal respiratory sounds.

In addition, sound event detection represents another notable application of audio-based sensing in Edge AI. Sound event detection systems leverage machine-learning algorithms to analyze audio signals and recognize specific sounds or events. By training on labeled data, these systems can identify various acoustic events, including glass breaking, gunshots, sirens, explosions, and more. This technology is essential for public safety and security applications, enabling real-time detection and response to potentially dangerous situations.

For instance, in smart cities, sound event detection can be integrated with surveillance systems to automatically identify and alert authorities of incidents like explosions, facilitating faster emergency response and enhancing overall security. Similarly, in industrial settings, these systems can detect abnormal sounds that indicate equipment failure, helping prevent accidents and improve operational safety. By harnessing the power of audio-based sensing and AI, sound event detection edge AI systems could efficiently contribute to developing intelligent and proactive environments, fostering more secure societies.



An audio-based sensing system.

Environmental Sensing Modalities

Environmental sensing modalities play a vital role in Edge AI systems by enabling devices to acquire and analyze data related to environmental conditions. Collecting and analyzing environmental data involves acquiring and processing sensor data, where environmental sensors generate continuous or periodic measurements that are logged and timestamped.
Data can be collected using wired or wireless communication protocols and transmitted for analysis to a central processing unit or a cloud server. Techniques such as statistical methods, machine learning algorithms, and anomaly detection are commonly employed to extract insights, identify patterns, and detect anomalies in environmental data.

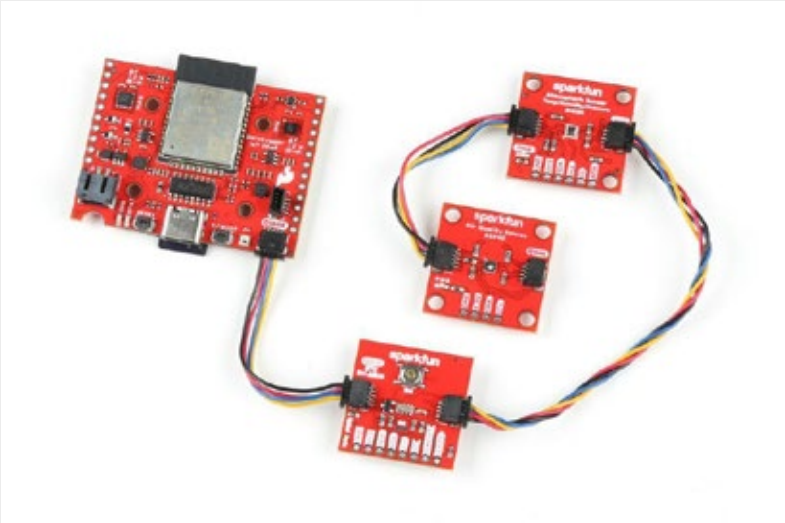
Environmental sensing modalities encompass diverse sensors designed to measure parameters such as temperature, humidity, and gas concentrations. These sensors play a pivotal role in Edge AI, offering a wide array of applications across various domains. In addition to contributing to climate control and energy management systems, environmental sensors serve as crucial safety measures, triggering alerts and facilitating proactive responses.

Data Collection Simplified: How Sparkfun Enables Data Logging

It's no secret that Edge computing requires accurate data collection. While SparkFun is not a software company, our aim is to develop a robust hardware ecosystem of Qwiic-compatible embedded sensors that enable Edge developers and partners to focus on innovating in this space. We want to see smarter sensors that enable users to collect and process data quickly for rapid development at the Edge.

As Kirk Benell, SparkFun CTO comments, "For over a decade, SparkFun has designed and built products aimed at simplifying the process of embedded device data logging, with the latest offerings delivering low-power, turnkey sensor connectivity to a growing set of SparkFun Qwiic boards."

A perfect example of how SparkFun simplifies the process of collecting data is showcased in our newest major product launch - the SparkFun DataLogger IoT - 9DoF. The DataLogger IoT is a data logger that comes preprogrammed to automatically log IMU, GPS, and various pressure, humidity, and distance sensors. It was specifically designed for users who just need to capture a lot of data to a CSV or JSON file and get back to their larger project. Save the data to a microSD card or send it wirelessly to your preferred Internet of Things (IoT) service!

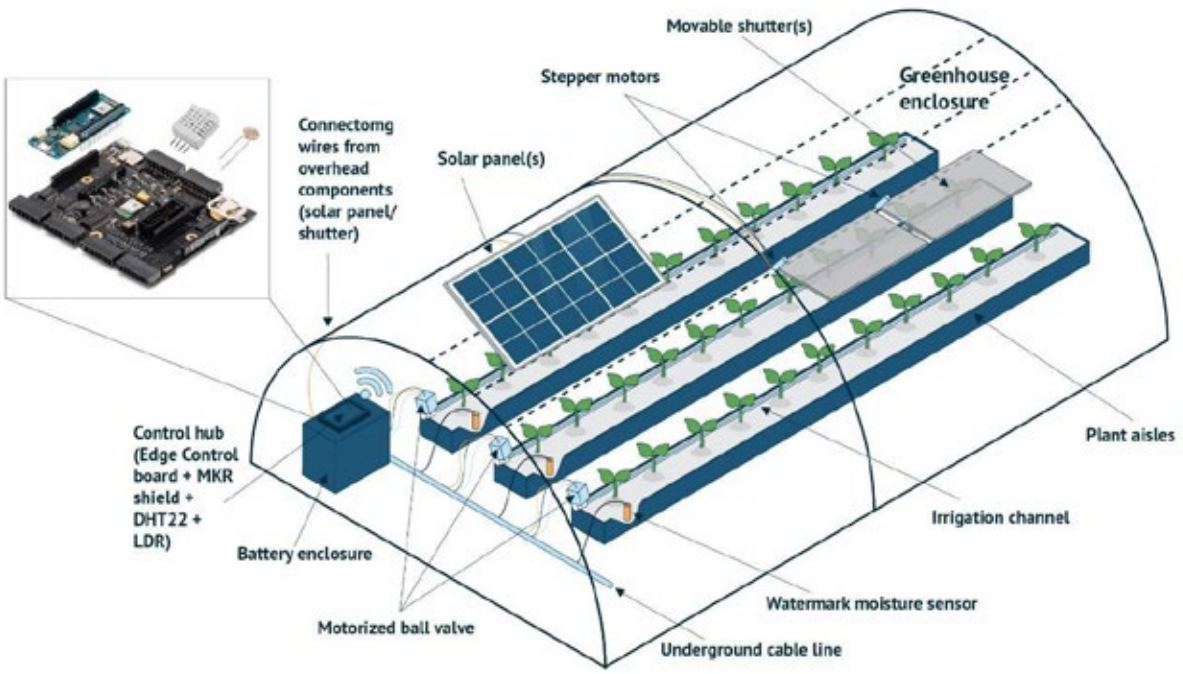


The SparkFun DataLogger IoT. Image credit: Sparkfun.

In industrial monitoring and workplace safety, environmental sensors monitor equipment, detect toxic or flammable gases, and ensure a safe work environment. In precision agriculture, they provide vital information for optimizing crop growth and disease prevention, enabling efficient irrigation management, early pest detection, and resource optimization.

One prominent example is the Arduino Edge Control by Arduino, which is used for controlling and monitoring remote environments while deploying AI on the edge. This board is suitable for a diverse range of applications requiring smart control in remote locations, such as agricultural automation, precision farming, and environmental and geophysical sensing. The Edge Control can enhance farming productivity by collecting data and making accurate data-driven decisions. It is especially useful in the control of highly sensitive farming processes such as hydroponics, aquaponics, and mushroom cultivation, which require precise temperature and humidity control. The components of the Edge Control board are designed with a higher temperature range to deal with harsh environments. It can be connected to a variety of sensors to collect real-time data on weather conditions, soil quality, CO_2 levels, pest infections, and crop growth.

Integrating environmental sensing into Edge AI systems makes real-time monitoring, predictive maintenance, and proactive safety measures possible, contributing to enhanced operational efficiency and worker well-being.



The Arduino Edge Control in an automated greenhouse application.

Other Sensing Modalities

In addition to vision-based, audio-based, and environmental sensing modalities, several other components play a meaningful role in data collection and analysis within Edge AI applications. These diverse modalities expand the capabilities of Edge AI systems by capturing valuable information from the physical world. Touch sensors, for instance, enable the detection and interpretation of tactile interactions, allowing devices to perceive touch inputs and provide haptic feedback. Capacitive touch sensors, resistive touch sensors, and surface acoustic wave sensors are commonly used in touchscreens and touch-sensitive surfaces, allowing for interactive and intuitive user interfaces. In Edge AI, touch sensors find applications in smart devices, human-machine interfaces, and interactive displays.

On the other hand, pressure sensors play a vital role in measuring the force applied to a surface, providing valuable insights into various systems and applications. In automotive settings, pressure sensors are employed for monitoring tire pressure, ensuring optimal vehicle performance and enhanced safety by detecting deviations from recommended pressure levels. In medical devices, pressure sensors measure fluid pressure accurately, enabling precise monitoring of blood pressure, respiratory functions, and intracranial pressure in healthcare settings. Additionally, in Heating, Ventilation, and Air Conditioning (HVAC) systems, pressure sensors are crucial for detecting changes in building pressure, allowing for efficient control and adjustment of ventilation systems to maintain optimal indoor air quality and energy efficiency. Integrating pressure sensors into Edge AI applications empowers real-time monitoring, data-driven decision-making, and predictive maintenance, leading to enhanced operational efficiency and improved user experience.

Finally, proximity sensors are highly versatile devices capable of detecting the presence or absence of objects in close proximity. They leverage various technologies, such as infrared, ultrasonic, and capacitive sensing, to achieve accurate detection. In the context of Edge AI applications, proximity sensors offer a multitude of possibilities. For instance, they can be employed in security systems to detect the presence of individuals or objects, enhancing overall surveillance effectiveness. In parking systems, proximity sensors play a key role in monitoring the proximity of vehicles, enabling efficient and automated parking assistance. Moreover, in touchless interfaces, proximity sensors are instrumental in detecting the approach of a hand or any other object, facilitating intuitive and hygienic user interactions.

The diverse range of sensing modalities discussed in this chapter offers unique characteristics that enable their suitability for specific applications in Edge AI. Combining sensing modalities with other types of sensors further enhances the capabilities of Edge AI applications, empowering them to provide intuitive and responsive interactions with the physical world. This integration opens up new possibilities for innovation and advancement in various fields, ranging from smart homes and healthcare to industrial automation and transportation.

Warehouse robotics use spatial AI that includes multiple sensing modalities to navigate between obstacles and carry cases around. Image credit: Gideon.ai.

Chapter VIII: Case Studies

Edge AI is making its way into various industries, leading to some surprising innovations. With enhanced data processing capabilities and relatively lightweight integration into existing systems, Edge AI is emerging as an optimal solution for gathering precise, accurate, and insightful information. In this chapter, we compiled several case studies that demonstrate some of the extraordinary innovations in Edge AI. But first, we asked industry experts at ST about their approach to Edge AI. Here's what they had to say.

Interview with ST:
A Glimpse into Edge AI From ST's Perspective

We sat down with Marc Dupaquier, Managing Director of the Artificial Intelligence Business Line at STMicroelectronics to discuss the Edge AI landscape and what role ST is playing to enable its development.

How would you describe Edge AI?

There are many definitions in the market for what Edge AI is, but for us, it is very simple, the Edge is where a signal becomes data. This means that the Edge is literally within the machine or the sensors; this is why we also speak about embedded AI, meaning Machine Learning that runs in a microcontroller (MCU) or microprocessor (MPU).

What is the adoption trajectory for machine learning at the Edge?

Machine Learning at the Edge is a rapidly growing segment of the market. There are about ten billion connected devices, of which a very small number have embedded intelligence. However, according to several analysts, 25% to 40% of them will have embedded algorithms in production within four years, which means that billions of devices will rapidly become embedded-AI smart. We observe this trend now, with hundreds of clients either already in production or testing prototypes for deployment in 2024. We believe that thanks to our 10-year investment in Edge AI, a very large number of these deployments will be on ST's platforms.

What are the benefits of Edge AI versus more traditional Cloud-based AI approaches?

Edge AI brings a set of advantages compared to the first generation of so-called smart devices, which were just connected to a Cloud and relied on the Cloud's intelligence.

1. An embedded device will not constantly stream data to the Cloud, which in itself brings a number of benefits, including smaller bandwidth requirements, lower cost of connection, and of course, improved security as you don't send raw data to the Cloud. Additionally, for battery-operated devices, the energy cost of constantly streaming data far exceeds the cost of inferring within the device and simply sending a report at a pre-set duration or when an anomaly occurs.

2. The total cost of operation will be way lower as there is simply no additional cost to inferring onto the device while running inferences in a Cloud can be costly, especially for vision-related use cases. This is why with the emergence of Edge AI, many solutions that simply were not affordable for the vendors as they required an expensive Cloud Service connection are now becoming affordable.
3. The total cost of energy of an Edge AI solution is significantly lower than a combo

"Smart Device/Cloud AI" approach, like Milliwatts vs. kilowatts, which, when multiplied by thousands of devices, creates a very important difference and something that we and many of our clients are now very sensitive to.

What is ST's machine learning at the edge's mission?

Our mission is to enable Edge AI by facilitating the creation, implementation, and operation of Edge AI solutions by providing a set of software, examples, and hardware that will create and run the embedded ML models.

What are the main design and implementation challenges that developers face, especially newcomers? And how do ST platforms address these challenges?

We are engaged with hundreds of customers implementing their projects and have observed a number of patterns:

First, data collection is very often hard to do for them, especially when they need to create abnormal data for predictive maintenance solutions. Secondly, they are often constrained by MCU size and RAM and flash limitations. Thirdly, they are not always sure of how to handle their approach or build their models.

Our solutions precisely address these issues. Cube.AI allows optimizing models so that they run smoothly on MCU platforms, and NanoEdge AI Studio has a fully integrated data capture and data preparation tool that allows for easy data input for model creation and an optimized search engine that will select the best model for our client's hardware constraints. And finally, with our model zoo, clients can have access to a wide range of examples that can inspire their own model creation.

What application segments is ST targeting with machine learning at the edge?

We think that Edge AI is a perfect fit for all kinds of needs where clients want a good level of service at an affordable price. Today, the two largest segments that our clients are deploying are predictive maintenance solutions and vision-related applications, where, in both cases, the availability of an edge solution is authorizing new deployments that clients had previously deselected because of the cost of ML Cloud processing.

Sensory Inc. - Revolutionizing User Experience with Voice-Activated AI Technologies

One of the most profound changes in computing came years ago in the then-unheralded world of human-machine interface (HMI) design. Innovators figured out how to marry the right hardware with the right software to create voice-activated applications. In the nearly three decades since this application caught fire, it's revolutionized everything from homes to personal assistants, computer programs, and automobiles. And leaders in the industry are just getting started.

Sensory Inc. has been developing software AI technologies for speech, sound, and vision out of Santa Clara, California, for the past 30 years. It has collaborated widely with industry leaders, including Arm and STMicroelectronics, to enhance the integration of its AI technologies and provide optimal solutions to customers. For example, Sensory has leveraged its long-standing partnership with Arm to support numerous chip partners using Arm Cortex-M4 microcontrollers with its Truly Hands-Free software, which enables devices with an extremely intelligent natural language user interface.

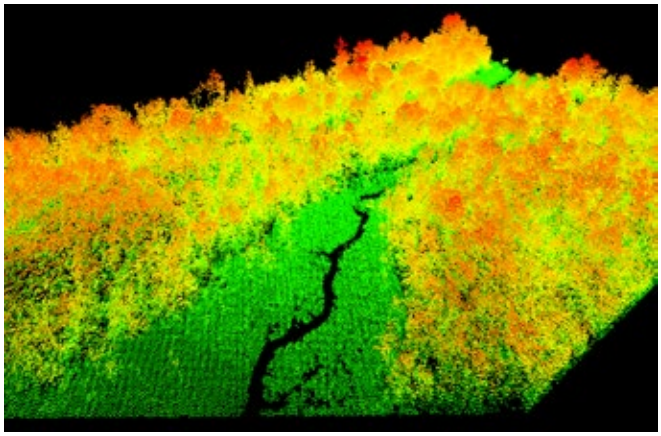
Two years ago, Sensory partnered with STMicroelectronics and their Arm Cortex-M7-based STM32H7 family. Sensory's TrulyNatural technology typically runs on an application process found in appliances, like microwave ovens, and enterprise types of products, like Zoom. It can provide natural language processing with more than 50,000 unique phrases that it recognizes. By leveraging the Cortex-M7-based STM32H7 family and integrating it with VoiceHub, Sensory's online portal enables developers to quickly create wake word models and voice control command sets for prototyping and proof-of-concept purposes.

TrulyNatural speech recognition technology from Sensory enables Smart Home applications powered by Arm. Image credit: Sensory.

Pachama - Predicting carbon capture in forests

Pachama is a San Francisco-based company with a mission to build market confidence in carbon credits with the help of Edge AI. The company uses remote sensing, including data from lidar, satellites, drones, and field plots, to estimate carbon captured by forests around the world and predict carbon storage in the future. Pachama has developed machine learning models using AI to process data captured by devices at the edge of a network.

Pachma relies on these sensors to monitor 'carbon sequestration,' the mechanism trees and other plants use to remove and store carbon dioxide from the atmosphere. Carbon sequestration is essential in mitigating greenhouse gasses that contribute to climate change. You can view global projects curated by Pachama in the Pachama Marketplace, a carbon credit marketplace built to assure companies and individuals that their carbon credit purchases are effective.



Pachama sensor imagery used to measure carbon captured by forests. Image credit: Pachama.

Activ Surgical - Real-time surgical visualizations

Activ Surgical is a medical startup developing surgical tools with Edge AI. The company aims to reduce complications during surgery and improve patient results. Activ Surgical's ActivSight Intelligent Light technology provides surgeons with real-time augmented imaging that enhances existing surgical techniques. With ActivSight, surgeons can view systems and functions humans can't normally see, like blood flow to anatomical structures.

ActivSight relies on the company's ActivEdge platform, a surgical imaging and guidance system that uses Edge AI. ActivEdge integrates with existing tools for surgical visualization, including human-operated scopes and surgical robotics. This integration benefits medical institutions because surgeons can perform operations with enhanced accuracy using standard equipment and techniques.

Activ Surgical's ActivSight imaging unit augments existing surgical equipment with advanced visualizations. Image credit: Active Surgical.

Medtronic - AI-Powered Endoscopy, Glucose Monitoring, and Cardiology

Medtronic is another company using Edge AI in medical applications. Medtronic has recently partnered with NVIDIA to implement the NVIDIA Holoscan platform in AI systems for medical devices. In particular, the Holoscan platform is integrated into Medtronic's GIGenius, an AI-assisted endoscopy module. With the help of this Edge AI technology, physicians can detect early signs of colorectal cancer.

Medtronic also develops other medical applications with edge AI. Its continuous glucose monitoring system measures glucose levels through a sensor inserted under the skin and uses AI to improve and personalize diabetes management. Medtronic's pacemaker uses sensing algorithms to detect and alert patients to cardiac arrhythmia. The company integrates edge AI into medical sensors to provide highly personalized medical care and detection.

GI Genius™ intelligent endoscopy module, which uses AI as an aid in detecting colorectal polyps during colonoscopy, potentially helping to prevent colorectal cancer (CRC). Image credit: Medtronic.

Fero Labs - Reducing Carbon Emissions with IoT

Fero Labs develops advanced software that optimizes industrial processes with Edge AI. Their process improvements have been shown to improve product quality, lower costs, and reduce manufacturing waste, with an average 35% reduction in CO2 emissions.

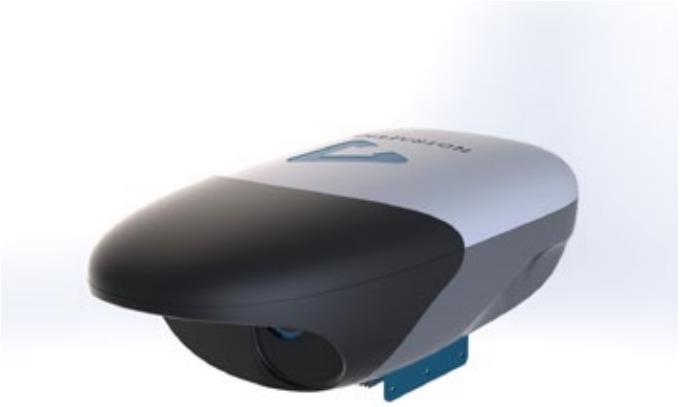
The company develops machine learning models designed to run on edge devices that are already present in factories. The software connects with standard industrial databases and processes data in its native format. With Fero Labs' software solutions, manufacturing plants can use existing equipment to predict product quality, identify problems with equipment, avoid flawed products, and improve process stability.

Factories can reduce CO2 emissions with machine learning software developed by Fero Labs. Image credit: Fero Labs.

NoTraffic - Traffic Management for Smart Cities

NoTraffic is an AI-based platform that manages traffic in real-time. NoTraffic's technology is another system built on NVIDIA technology, combining sensors and software to process and quickly respond to traffic conditions. NoTraffic's Edge AI technology detects, identifies, and tracks road users, including their speed and direction as they approach an intersection, and adjusts the traffic signal accordingly.

NoTraffic has gained worldwide attention and has partnerships with US state transportation departments in California and Arizona. The platform has been shown to reduce traffic wait times by as much as 50%, which may also contribute to reduced vehicle emissions. An added benefit of NoTraffic's technology is that smart cities with traffic systems built on Edge AI can quickly implement policies and analyze data from traffic flow.



NoTraffic IoT sensor that combines camera and radar to detect road users and predict traffic flow. Image credit: NoTraffic.

BloomX - Pollination with Robot Biomimicry

BloomX has developed a surprising application for Edge AI by mimicking the work of natural plant pollinators using predictive algorithms and purpose-built robotics. BloomX's line of EV pollinators includes the Robee robot, which uses mechanical arms that have undergone years of testing and refinement to mimic the vibration that the bumblebee provides. In the natural world, this vibration is a mechanism that releases pollen from flowering plants. To maximize crop growth, pollinator robots rely on Edge AI to pinpoint the precise window for pollination, with onboard equipment that uses GPS-tracking and environmental sensors, giving growers real-time tracking data and controls.

BloomX indicates that their technology may result in up to a 30% increase in crop yield. The company's clients include growers around the world, particularly blueberry and avocado growers in Latin America, South America, Africa, and the US. The company recently secured $8 million in funding from global organizations and manufacturers interested in developing this technology.

The BloomX Robee mimics natural pollination methods with robotic end effectors and AI sensors. Image credit: BloomX.

Starkey - Advanced Performance for Hearing Aids

Starkey is another company using Edge AI to bring innovation to the medical industry. The company's hearing aid products, the Livio Edge AI and Evolv AI, feature Starkey's Edge Mode technology, which can sense and respond to the wearer's sound environment. Edge Mode uses machine learning algorithms to conduct real-time analysis of the listening environment and make adjustments to gain, noise management, and speech audibility.

Starkey's hearing aids are a consumer-ready edge AI product available on the market today. However, Starkey's most advanced hearing aids with Edge Mode may cost USD 4,200 or more, depending on features. Their Edge Mode hearing aids are rechargeable and offer fall detection, user-friendly control, customization via a smartphone app, and other features. According to a review in Forbes Health, these devices may mitigate the issues associated with hearing loss, such as isolation and depression, that result when users have difficulty adjusting to traditional hearing aids.

Livio Edge AI, a hearing aid that uses sensors to detect and respond to the wearer's sound environment. Image credit: Starkey.

Motional - Autonomous Robotaxis

A report on Edge AI use cases would be incomplete without a nod to autonomous vehicles, an industry taking a major role in AI and hardware innovations. Motional is one such company in this sector that is implementing Edge AI in autonomous vehicles. The company is building autonomous robotaxis that use over 30 sensors with 360-degree vision and onboard AI computing. Motional aims to develop AI-controlled vehicles that adhere to the highest safety standards and integrate easily into existing ridesharing networks. Motional's flagship, the IONIQ 5, was designed with an AI-first approach, meaning that its systems are fully engineered for driverless operation.

Motional is a joint venture between the automotive tech company Aptiv and the automotive manufacturer Hyundai. Recently, Motional partnered with Uber to launch an autonomous ride-hailing service in Las Vegas. Currently, the project employs human operators, but the companies aim for completely driverless vehicles in 2023. In Santa Monica, Motional partnered with Uber Eats to make autonomous food deliveries using Motional vehicles. The company also conducts testing in Boston, Pittsburgh, and Singapore, gathering data from diverse environments for improved onboard algorithms.

Edge AI is a versatile technology with diverse applications. It is a scalable and adaptable tool that has the potential to transform every industry. The ability to use existing equipment is a common goal among companies implementing Edge AI—a factor that lowers the barrier to adopting this beneficial technology. Another common feature of Edge AI technology is that it offers real-time insight, meaning that data can be immediately applied in critical situations, which could be anything from a surgical procedure to monitoring factory equipment or to a quick response to changing traffic conditions. Edge AI complements, enhances, and brings innovation to data processing capabilities.

Motional's IONIQ 5 robotaxi, an autonomous vehicle equipped with sensors and onboard AI computing to keep passengers, pedestrians, and surrounding vehicles safe. Image credit: Motional.

Chapter IX: Challenges of Edge AI

Introducing Edge AI has brought numerous advantages to the modern world of computing. With the Edge AI philosophy, data processing can be performed locally on a device, eliminating the need for sending large amounts of data to a centralized server.

This not only reduces data transfer requirements but also results in faster response times and increased reliability, as the device can continue to operate even if it loses connection with the network. Edge technology also provides enhanced data privacy, as sensitive data can be processed locally and does not need to be transmitted to a third-party server.

Despite the numerous potential benefits of Edge AI, there are also several challenges associated with its implementation and usability. In this chapter, we will examine these challenges in detail, explaining how they can impact the efficiency, performance, and overall success of Edge AI deployments.

Data Management Challenges

Data management poses a significant challenge in Edge AI, as edge devices often capture real-time data that can be incomplete or noisy, leading to inaccurate predictions and poor performance. In fact, the challenges in computing at the edge are related to data movement, which has an impact on power, efficiency, latency, and real-time decision-making. Data movement is influenced by the volume and velocity of data, which need to be orchestrated from the data center to the endpoint. The volume of data has security implications, power implications, computation requirements, and application-level implications for real-time decision-making. The concept of the edge aims to reduce data movement, reduce latency, and enable more real-time decision-making through distributed intelligence.

“The more decision-making an endpoint can make without consulting the data center, the more real-time it could be.”

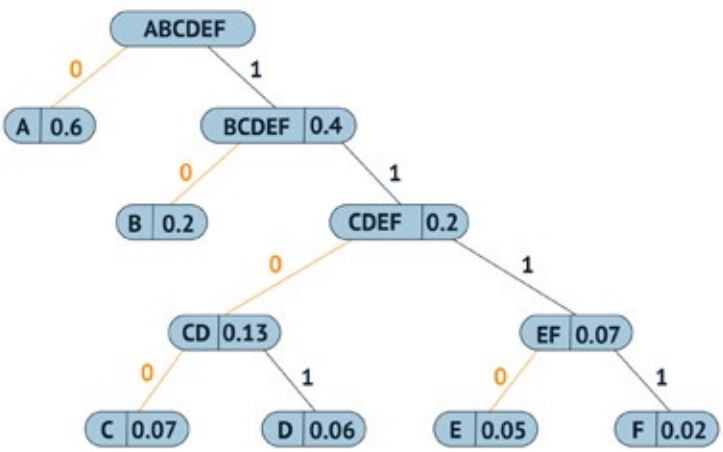
Chowdary Yanamadala, Senior Director, Technology Strategy, Arm’s Advanced Technology Group

Researchers have proposed various solutions to improve the quality of data collected by Edge AI devices. One promising approach is federated learning, which involves training AI models on data distributed across multiple edge devices, enhancing the quality and diversity of training data.

Another data management issue is the limited storage capacity of edge devices. Since Edge AI devices typically have limited memory storage, storing larger quantities of data produced by edge applications can be challenging. To overcome this issue, data compression techniques such as Huffman coding and Lempel-Ziv-Welch (LZW) compression can be employed to reduce the size of data stored on edge devices, optimizing memory usage.

When discussing data management practices, data governance is another crucial aspect that is often overlooked in the context of Edge AI devices. As AI applications become more prevalent, it becomes increasingly important to ensure that data generated by edge devices are appropriately managed and aligned with regulations and policies. However, governing data on edge devices can be challenging, particularly in complex enterprise environments.

To overcome this obstacle, blockchain-based data governance frameworks such as Hyperledger Fabric and Ethereum have emerged as potential solutions. These frameworks enable secure and transparent data management and provide a distributed and decentralized approach to managing and tracking data. By utilizing blockchain technology, data governance on Edge devices has shown to be more efficient, secure, and reliable, thus ensuring compliance with required regulations.



A Huffman coding tree built from the following character frequency table: A=0.6, B=0.2, C=0.07, D=0.06, E=0.05, F=0.02. Adapted from Manning by Wevolver.

Integration Challenges

Integrating Edge AI with other systems can be challenging for many reasons, where compatibility issues dominate. In general, incompatibility issues can arise from differences in hardware, software, and communication protocols. Each edge device can have unique specifications, architecture, and interfaces, making integration with other devices and systems difficult and not routine work. At the same time, AI models may require specific software libraries, frameworks, and programming languages that may not be compatible with other systems.

To overcome these challenges, the open-source community has contributed integration frameworks, proposed new industry standards, and develoWped efficient communication protocols to solve interoperability issues and promote standardization. For example, the Open Neural Network Exchange (ONNX) is an open-source approach applicable to deep learning models that can be used across different frameworks and platforms, promoting interoperability and reducing compatibility issues.

Another interesting example is how Synaptics tackles integration challenges. They tie wireless and edge together by providing chips for several wireless techs for the IoT, such as Wi-Fi, Bluetooth, IEEE 802.15.4 Thread/Zigbee, and Matter. They also provide ultra-low energy (ULE) technology, which operates in an interference-free band (1.9 GHz) and supports secure, wide-coverage wireless communications with support for two-way voice and data. Venkat Kodavati, Senior Vice President and General Manager of Wireless Products at Synaptics, stated that "When meta or other sensor data need to be transported to an aggregation point for further processing, it's critical that the appropriate wireless technology be used to ensure a robust connection with the lowest possible latency and power consumption."

Synaptics is also working with laptop manufacturers on the "Integration of visual sensing based on ultra-low-power machine learning (ML) algorithms," as Kodavati explained. "These algorithms analyze user engagement and detect on-lookers to save power and enhance security on the laptop itself." He added, "The algorithms are being optimized for efficiency and compatibility with any system hardware. The efficiency of Synaptics' own Edge AI visual sensing hardware, which began with Katana, is also improving rapidly to adapt to upcoming applications that will incorporate more sensing capabilities as we shift to more advanced context-aware computing at the edge."

Security Challenges

In addition to compatibility issues, integrating Edge AI with other systems may be followed with significant security risks. Edge devices often collect and process sensitive data, including personal health records, financial information, and biometric data, which makes them vulnerable to security threats such as data breaches, cyber-attacks, and privacy violations.

"The biggest challenge for security is consistency, which also applies to software in general."

David Maidment, Senior Director, Secure Device Ecosystem, Arm

Therefore, security measures should be considered from the initial design phase of Edge AI systems. Techniques such as secure boot and hardware root of trust (RoT) can ensure the integrity of edge devices. Similarly, secure software development practices like threat modeling and code reviews can prevent common vulnerabilities. "Standards such as Arm SystemReady play a role in addressing this challenge by providing a standardized OS installation and boot for edge devices," Maidment said. "SystemReady includes secure boot and secure capture updates to ensure consistency in the way the operating system lands on the platform and remains secure. This is important for reducing fragmentation and lowering the cost of ownership over the device's lifetime."

Furthermore, integration with cloud-based services may also introduce additional security risks, which can be prevented by exploiting techniques such as data encryption, secure authentication, and secure communication protocols. From the perspective of AI models, techniques such as federated learning, differential privacy, and homomorphic encryption have been used to train AI models on sensitive data without compromising privacy. Moreover, anomaly detection techniques, such as intrusion detection systems, can be used to detect and mitigate attacks targeting Edge AI systems, ensuring the entire system's security.

"The move to Edge AI processing focuses on reducing or eliminating raw data sent to the cloud for processing. This enhances the security of the system or service," explains Nandan Nayampally. An on-device learning capability, like the one offered by BrainChip, not only allows for customization and personalization untethered from the cloud, but it also "stores this as weights rather than data, further enhancing privacy and security, while still allowing this learning to be shared by authorized users when needed."

Latency Challenges

Latency is a significant issue that can significantly affect the performance of Edge AI systems. These systems face three types of latency challenges: input latency, processing latency, and output latency.

Input latency is the delay between the time a data sample is captured by an edge device and the time it is processed by an AI model. It can result from factors such as slow sensor response time, data transmission delay, and data pre-processing overhead. Input latency can impact the accuracy and timeliness of AI predictions, leading to missed opportunities for real-time decision-making.

Processing latency, on the other hand, refers to the delay between the time an AI model receives a data sample and the time it generates a prediction. Factors such as the complexity of the AI model, the size of the input data, and the processing power of the edge device can cause processing latency. It can affect the real-time responsiveness of AI predictions and may cause delays in critical applications such as medical diagnosis and autonomous driving.

Output latency is the delay between the time an AI prediction is generated and the time it is transmitted to the user or downstream system. Various factors, such as network congestion, communication protocol overhead, and device-to-device synchronization, can cause it. Output latency can impact the usability and effectiveness of AI predictions and may cause delays in decision-making and action-taking.

To address latency challenges, various techniques such as edge caching, edge computing, and federated learning are widely used in Edge AI systems. Edge caching helps reduce input latency by storing frequently accessed data closer to the edge of the network. This technique can store pre-trained AI models, reference data, and other relevant data, thus improving the real-time responsiveness of AI predictions.

Additionally, edge computing reduces processing latency by moving AI processing from the cloud to the edge of the network. By deploying lightweight AI models such as decision trees and rule-based systems on edge devices with limited processing power and storage capacity, it can eliminate the need for data transmission, thus improving the real-time responsiveness of AI predictions. Another notable approach is federated learning, which allows distributed AI model training on edge devices without transferring raw data to the cloud. This technique reduces the risk of data privacy violations and lowers output latency. By training AI models on diverse data sources such as mobile devices and IoT sensors, federated learning can improve the accuracy and generalization of AI predictions.

Scalability Challenges

Edge AI systems can face significant scalability challenges that can affect their performance, reliability, and flexibility. Scalability refers to the ability of a system to handle increasing amounts of data, users, or devices without compromising efficiency. These challenges can be classified into three categories: computational scalability, data scalability, and system scalability.

Computational scalability is the ability of an Edge AI system to process increasing amounts of data without exceeding the processing power and storage capacity of edge devices. Edge devices' limited processing power, memory, and storage can restrict the size and complexity of AI models and hinder their accuracy and responsiveness.

Data scalability, on the other hand, is the capability of an Edge AI system to handle increasing amounts of data without compromising performance. Processing large amounts of data in real-time on edge devices can be difficult due to their limited data transfer capacity and unreliable connectivity, which may restrict the quantity and quality of data that can be transmitted and processed.

System scalability refers to the capacity of an Edge AI system to manage growing numbers of edge devices and users efficiently and appropriately. This can be difficult because Edge AI systems require distributed processing and coordination, which can introduce latency, overhead, and complexity.

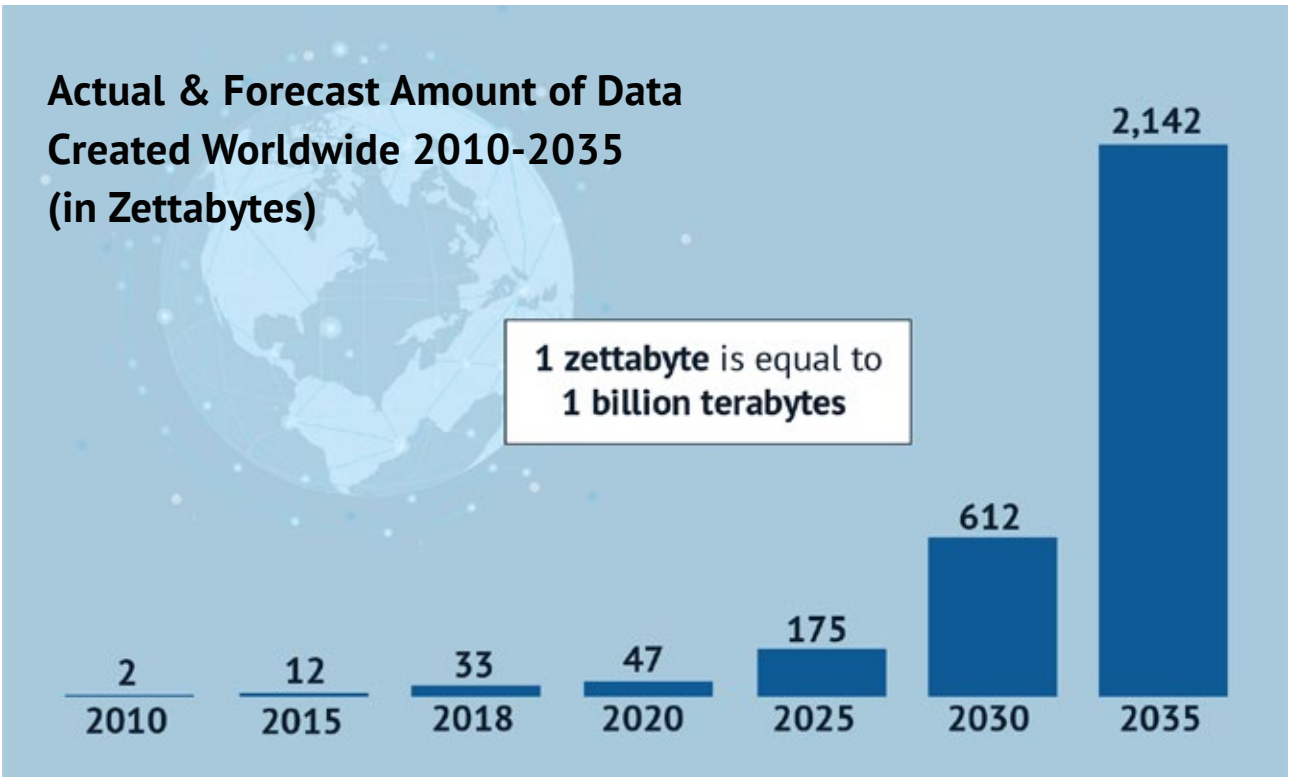
To address these challenges, techniques such as load balancing, parallel processing, and distributed computing can be used to optimize system scalability and enhance the overall performance and reliability of Edge AI systems. By leveraging these techniques, Edge AI systems can achieve the scalability required to handle the increasing volume and diversity of data generated by edge devices, users, and applications.

Several potential solutions have been introduced to address scalability issues in Edge AI systems, including edge orchestration, edge-to-cloud coordination, and network slicing. Edge orchestration involves managing the deployment and coordination of AI models on multiple edge devices, enabling distributed processing and load balancing. This approach optimizes the allocation of processing resources, such as CPU and GPU, and minimizes communication overhead between edge devices without being hindered by the limited processing power, memory, and

storage constraints of the edge devices. Another solution is Edge-to-cloud coordination, which integrates Edge AI systems with cloud-based AI systems, enabling hybrid processing and seamless data exchange. It allows heavy processing tasks, such as AI model training and validation, to be offloaded to the cloud and enables data aggregation and analysis from multiple edge devices. Finally, Network slicing can address scalability issues by partitioning wireless networks into multiple virtual networks, each with a unique set of resources and quality-of-service guarantees. This approach can be used to allocate network resources based on Edge AI systems' specific requirements and priorities, such as low latency, wide data transfer capacity, and high reliability, without being limited by the challenges posed by restricted data transfer capacity and unreliable connectivity.

In addition to these challenges, one more challenge requires mentioning, and that is the AI modeling challenge. As explained by Shay Kamin Braun of Synaptics, "When you design an edge device model before deployment, you try to test it in as many environments as possible so the model does what it's supposed to do without mistakes." For example, a device that is supposed to detect a person may face two separate issues: (1) not detecting a person that is there or (2) misdetecting something else for a person.

It is quite difficult to scale testing before deployment, especially for an edge model running inside multiple devices and getting inputs from multiple sensors, which means there remain untested environments. Continuous learning after the models have been deployed is not easy, particularly for battery-powered devices. That is why companies that are able to scale their testing and mature their models prior to deployment are the ones expected to win.



A forecast of the amount of data created worldwide between 2010 and 2035. Adapted from Statista by Wevolver.

Cost Challenges

Edge AI systems can easily face cost challenges that constrain their adoption and impact. First on the list are hardware costs, which refer to the cost of specialized edge hardware, sensors, and actuators. Edge devices often require specific CPU and GPU hardware components that can accelerate AI workloads and improve performance while keeping their size minimalistic.

On the other hand, software costs refer to the cost of developing, deploying, and maintaining software applications and algorithms required to support Edge AI applications. Developing AI applications and algorithms can be time-consuming and require specialized skills and knowledge, which can increase development costs. Furthermore, deploying and maintaining software applications and algorithms on Edge AI devices can be challenging due to edge devices' limited processing power and storage capacity, which can increase operational costs.

Edge AI systems may also face several costs that can impact their adoption and effectiveness beyond hardware, software, and operational costs. For example, data storage and management costs are a significant concern, with edge devices generating and storing vast amounts of data that can quickly fill up limited storage capacity. Also, reliable and high-speed network connectivity is essential for edge devices, but setting up and maintaining robust network infrastructure can be expensive, particularly in remote or rural areas.

Computing costs are a critical challenge that can impact the decision-making process for businesses. "By being power efficient, we can indirectly contribute to lowering the cost of operations," Yanamadala said. "If our architecture is 30% more power-efficient than others but – for argument's sake – costs 10% more to integrate, we can argue that the additional upfront investment is worth it because it leads to continued benefits of lower cost of operations. This trade-off between capital expenditure and operating expenditure is a common decision-making process for enterprises when moving to the cloud or data center side of things."

Data privacy and security are also critical aspects, as edge devices collect sensitive data that must be protected from breaches. Furthermore, Edge AI applications often require a specialized network infrastructure, such as high-speed networks with minimal delay or latency, that can support real-time processing and analysis, adding to infrastructure costs. To reduce these costs, organizations can leverage existing network infrastructure, implement edge-to-cloud coordination, or adopt open-source software.

> "One of the hidden costs in edge computing is the cost of computing."
>
> **Chowdary Yanamadala,** Senior Director, Technology Strategy, Arm's Advanced Technology Group

Power Consumption Challenges

Edge AI systems can also face challenges related to high energy consumption, which can limit their adoption and impact, especially in remote and harsh environments such as industrial plants, agricultural fields, and highways. High energy requirements can be attributed to the need for powerful computing resources to process and analyze data in real-time. Edge devices often require high-performance processors, memory, and storage devices that consume significant amounts of energy, making it challenging to power such systems in energy-constrained environments. Additionally, always-on connectivity and data transfer between Edge devices and the cloud can further increase energy consumption.

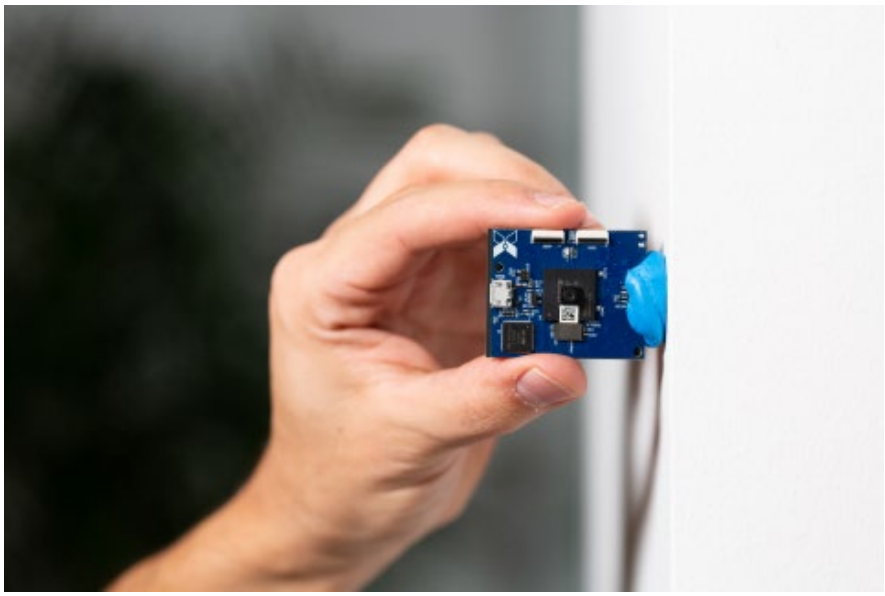
To address these challenges, one solution is to use energy-efficient hardware, such as low-power processors and memory, to reduce energy consumption without compromising performance. This can be achieved using edge computing hardware designed explicitly for Edge AI applications. For example, some edge computing devices are equipped with ARM-based processors optimized for low-power consumption while delivering high performance. Another solution is to optimize the software algorithms used in Edge AI systems. Techniques such as reducing unnecessary data transmissions and improving the accuracy of predictive models can decrease the overall computational load, resulting in lower energy consumption.

In addition, integrating renewable energy sources such as solar and wind power with energy-efficient hardware can enable Edge devices to operate autonomously, reducing their energy consumption and carbon footprint. For instance, edge devices can be equipped with solar panels or wind turbines to generate electricity locally. Battery storage systems can also be used to store excess energy generated during peak times, allowing devices to operate during periods of low energy availability. By leveraging renewable energy sources, Edge AI systems can operate sustainably and reduce their environmental impact.

Potential Solutions

Implementing artificial intelligence at the edge presents an intricate set of challenges, especially when operating within the constraints of an edge device such as a smartphone. Here, the AI accelerator's footprint is necessarily small, and its energy consumption must be meticulously managed. These strict requirements, however, inevitably lead to a trade-off, as the limited physical space and energy restrict storage and computational capacities.
To alleviate such concerns, chip designers may transition to smaller nodes (e.g., from a 12nm to a 5nm process) and modify the chip's architecture (via 3D stacking, for example). While these hardware-oriented approaches effectively miniaturize the physical constraints, they often fall short when grappling with the inherent limitations of Edge AI applications. In this realm, massive computing, small form factors, energy efficiency, and cost-effectiveness must go hand in hand, a challenging balance to strike.

Confronting these challenges, Edge AI demands an innovative approach: co-designing hardware and software. This paradigm encompasses tactics of data compression, like data quantization and network pruning. By compressing the data, the chip requires less storage and memory and can rely on simpler computational units - those operating on fewer bits, for instance - to execute the AI algorithms.



A standalone, battery-free, solar-powered AI technology from Xnor.Ai that will enable AI to run in an always-on mode on a range of edge devices.
Image credit: Xnor.Ai.

This reduction in memory and computational complexity reciprocally enhances energy efficiency - a classic case of killing multiple birds with one stone.

However, data compression is not without its own drawbacks. A significant concern is that it may precipitate an undesirable loss in the AI model's accuracy. To circumvent this, algorithm developers must carefully identify which segments of the AI network can withstand compression and to what extent without leading to significant degradation of accuracy.

Moreover, to ensure widespread usability, the compression algorithm should be optimized for speedy execution and compatibility with a broad range of networks. Concurrently, hardware developers should focus on devising ways to maximize the efficiency of compressed data usage while communicating the potential and limitations of their implementations to the algorithm developers.

In conclusion, the successful deployment of AI at the edge requires a harmonious blend of hardware and software co-design, informed by a deep understanding of the constraints and potentials of both domains.

"Balancing the quest for miniaturization, energy efficiency, cost-effectiveness, and model accuracy is no simple task. Still, with careful coordination and communication, we can pave the way for the next wave of AI breakthroughs at the edge."

Dr. Bram Verhoef, Head of Machine Learning at Axelera AI

Chapter X: The Future of Edge AI

In the coming years, the increasing demand for real-time, low-latency, and privacy-friendly AI applications will lead to a proliferation of Edge AI deployments. The latter will be continually more accurate and efficient, leveraging technological advances in communication networks (e.g., 5G and 6G networking), artificial intelligence (e.g., neuromorphic computing, data-efficient AI), and IoT devices. In this direction, the role of researchers and the open-source communities will be fundamental in driving the evolution of edge AI.

The development and deployment of efficient edge AI applications hinge on the evolution of different technologies destined to support edge devices and compute-efficient AI functions. Specifically, the following technological evolutions will shape the future of edge AI.

Emergence and Rise of 5G/6G Networks

Deploying real-time, low-latency Edge AI applications requires ultra-fast connectivity and high bandwidth availability. This is why conventional 4G and LTE (Long Term Evolution) networks are not entirely suitable for delivering Edge AI benefits at scale. During the last couple of years, the advent of 5G technology has opened new opportunities for high-performance Edge AI deployments.

5G networks deliver thousands of times more bandwidth than previous generation networks while at the same time supporting edge computing applications in environments where 1000s of IoT devices are deployed. As such, 5G networks facilitate Edge AI applications to access large amounts of training data and to operate without disruptions in crowded environments and device-saturated contexts. As 5G networks continue to expand, we can expect a proliferation of AI-enabled intelligent edge devices that will perform complex tasks and make autonomous decisions in real time.

In the future, 6G networks will also emerge to offer even faster speeds. 6G infrastructures are currently under research and development. They are expected to become commercially available after 2030 and use higher frequency bands than 5G (e.g., they will operate in the 30 to 300 GHz millimeter waves). 6G networks will increase bandwidth availability and network reliability, which will be vital in supporting future large-scale Edge AI applications. The latter will include numerous heterogeneous AI-based devices running a multitude of AI algorithms, including algorithms that will operate based on many data points thanks to 6G's capacity.

5G towers of a mobile telephony observatory in Paris.

Neuromorphic Computing: Increase AI Intelligence by Mimicking the Human Brain

The future of Edge AI will also be shaped by novel AI paradigms such as neuromorphic computing. This approach mimics the human brain's structure and functionality by emulating the neural networks and synaptic connections in our brains. It is based on novel neuromorphic chips that process information more efficiently while adapting faster and more effectively to new situations. In practice, neuromorphic chips comprise many artificial neurons and artificial synapses that can mimic the functioning of brain spikes. Therefore, neuromorphic computing research brings us a step closer to understanding, decoding, and exploiting the code of the human brain in AI applications.

Neuromorphic computing chips are well-suited to deliver Edge AI benefits at scale. This is because they consume less power and provide faster processing speeds than conventional processors. Most importantly, they are equipping Edge AI systems with human-brain-like reasoning capabilities that will be extremely useful in many pervasive applications (e.g., obstacle avoidance, robust acoustic perception, etc.). As neuromorphic computing technology matures, it will enable a new generation of AI-based edge devices that can learn and adapt in real-time.

Event-based Processing & Learning: BrainChip's Neuromorphic AI Solution

BrainChip is one of the pioneers of bringing neuromorphic computing to the edge. While traditional neuromorphic approaches have used analog designs to mimic the neuron and synapse, BrainChip has taken a novel approach on three counts.

- Firstly, not only do they support spiking neural nets, but they have applied event-based execution to traditional convolutional networks, thereby rendering neuromorphic computing mainstream today. This allows current CNN/RNN models to run much more efficiently and drives far more capable performance on extremely low-footprint, low-power devices at the sensor.

- Secondly, their design is a fully digital design that is portable and reliable.

- Thirdly, delivering on-device learning allows for personalization, customization, and other learning untethered from the cloud.

Brainchip's Akida neural processor is offered as IP and is configurable from energy-harvesting applications at the sensor edge to high-performance yet power-efficient solutions at the network edge. It is sensor-agnostic and has been demonstrated on a variety of sensors.

As a self-managed neural processor that executes most networks completely in hardware without CPU intervention, it addresses key congestion and system bandwidth challenges in embedded SoCs while delivering highly efficient performance. With support for INT8 down to INT1 and skip connections, it handles most complex networks today, along with spiking neural nets.

This led NASA to select BrainChip's first silicon platform in 2021 to demonstrate in-space autonomy and cognition in one of the most extreme power- and thermaly-constrained applications. Similarly, Mercedes Benz demonstrated BrainChip in their EQXX concept vehicle that can go over 1000 km on a single charge.

In the latest generation, Brainchip has taken another big step of adding Temporal Event Based Neural Nets (TENNs) and complementary separable 3D convolutions that speed up some complex time-series data applications by 500x while radically reducing model size and footprint, but without compromising accuracy. This enables a new class of compact, cost-effective devices to support high-res video object detection, security/surveillance, audio, health, and industrial applications.

While neuromorphic computing is still discussed as a future paradigm, BrainChip is already bringing this paradigm to market.

Data-Efficient AI: Maximizing Value in the Absence of Adequate Quality Data

One of the significant challenges in AI is the need for vast amounts of data to train machine learning algorithms (e.g., deep learning) effectively. There is frequently a lack of adequate quality data to train such algorithms. This issue is prevalent in the case of Edge AI systems and applications, given the specialized nature of embedded machine learning and TinyML systems, which require their own data collection processes. Moreover, Edge AI systems face computational and storage constraints that prevent them from fully leveraging huge AI models and large numbers of data points.

Data-efficient AI techniques aim at overcoming the above-listed limitations by enabling AI models to learn from limited data samples. Thus, they obviate the need for large-scale data collection while lowering the computational requirements of Edge AI systems.

Ongoing research on data-efficient AI explores many techniques, ranging from augmenting and using pre-trained models with domain knowledge (e.g., transfer learning) to paradigms that engage humans in the data labeling processes (e.g., active learning) as part of human-AI interactions. There are also popular data-efficient techniques that reduce the size of the AI model (e.g., model pruning) to enable space-efficient models that can fit on edge devices. Moreover, some techniques are inherently capable of learning from limited data samples, such as one-shot learning and few-shot learning.

Overall, future data-efficient AI methods will help edge devices make accurate predictions and decisions with minimal data input. Data-efficient techniques will also boost the overall performance of Edge AI systems while at the same time reducing the time and resources required for model training.

In-Memory Computing for Edge AI

In-memory computing is one more technological trend that will impact the future of Edge AI. It is about storing and processing data directly within the memory of a device rather than relying on traditional storage systems (e.g., disk). This approach is set to reduce data access times significantly and accelerate Edge AI systems' computational speeds. Therefore, it will further boost Edge AI systems' real-time analytics and decision-making capabilities.

In the future, Edge AI applications will have to process increased volumes of data rapidly. Hence, in-memory computing will become increasingly important for optimizing Edge AI's performance and efficiency. In particular, it will enable Edge AI devices to process complex algorithms and extract valuable insights from data at unparalleled speeds.

Digital In-Memory Computing: An Axelera AI Solution

We asked Evangelos Eleftheriou, CTO and Co-founder at Axelera AI, about his thoughts on in-memory computing. Here's what he had to say.

The latency that arises from the increasing gap between the speed of memory and processing units, often referred to as the memory wall, is an example of a critical performance issue for various AI tasks. In the same way, the energy required to move data is another significant challenge for computing systems, particularly those that have strict power limitations due to cooling constraints, as well as for the wide range of battery-powered mobile devices. Therefore, new computing architectures that better integrate memory and processing are needed.

Near-memory computing, a potential solution, reduces the physical distance and time to access memory. It benefits from advancements in die stacking and technologies like high memory cube (HMC) and high bandwidth memory (HBM).

In-memory computing (IMC) is a fundamentally different approach to data processing, in which specific computations are carried out directly within the memory by arranging the memory in crossbar arrays. IMC units can tackle latency and energy problems and also improve computational time complexity due to the high level of parallelism achieved through the dense array of memory devices that perform computations simultaneously.

Crossbar arrays of such memory devices can be used to store a matrix and perform matrix-vector multiplications (MVMs) without intermediate movement of data. This efficiency is especially beneficial for training and inference in deep neural networks, where energy efficiency is critical. As 70-90% of deep learning operations are matrix-vector multiplications, applications with many AI components, such as computer vision and natural language processing, can benefit from this technology. The efficient matrix-vector multiplication via in-memory computing is very attractive for the training and inference of deep neural networks, particularly for inference applications such as computer vision and natural language processing, where high energy efficiency is critical. In fact, matrix-vector multiplications constitute 70-90% of all deep learning operations.

There are two classes of memory devices:

- The conventional class, including dynamic random-access memory (DRAM), static random-access memory (SRAM), and Flash memory
- An emerging class of resistive memory devices known as memristors

Traditional and emerging memory technologies can perform a range of in-memory logic and arithmetic operations, as well as MVM operations.

SRAM, having the fastest read and write time and highest endurance, enables high-performance in-memory computing for both inference and training applications. It follows the scaling of CMOS technology and requires standard materials and processes. Its main drawbacks are its volatility and larger cell size.

Axelera AI introduced an SRAM-based digital in-memory computing (D-IMC) engine, which is immune to noise and memory non-idealities that affect the precision of analog in-memory computing. D-IMC supports INT8 activations and weights with INT32 accumulation, maintaining full precision for various applications without retraining. The D-IMC engine of the matrix-vector-multiplier is a handcrafted, full-custom design that interleaves the weight storage and the compute units in an extremely dense fashion, thus reducing energy consumption while maintaining high energy efficiency even at low utilization.

Axelera AI's software stack with a Digital In-Memory Computing (D-IMC) engine. Image credit: EENews Europe.

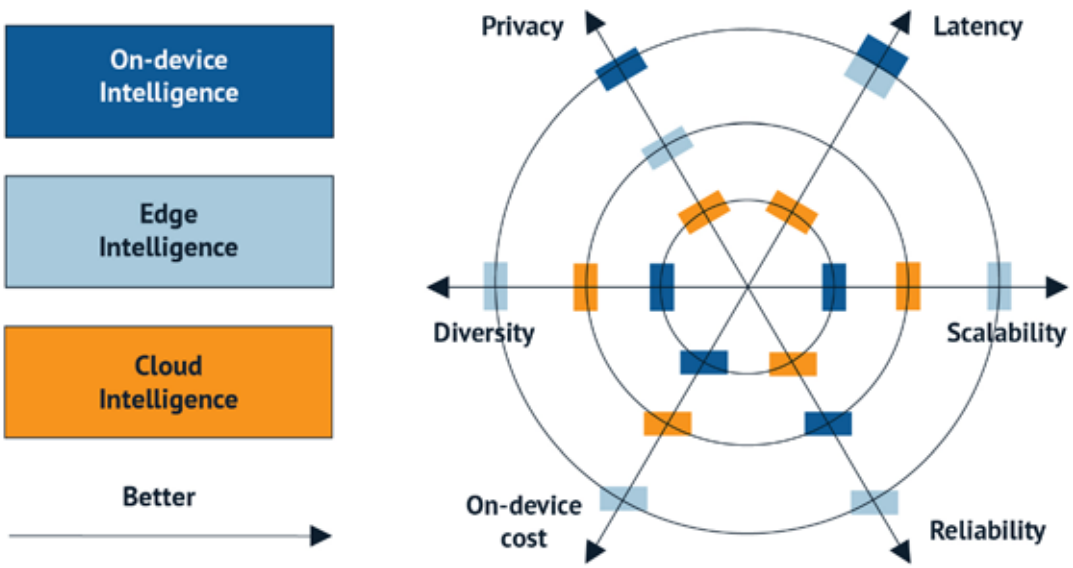
Distributed Learning Paradigms

Earlier parts of the report have outlined Edge AI approaches based on decentralized learning paradigms like federated learning. Federated learning enables edge devices to collaborate toward training a 'global' shared machine learning model from 'local' models without exchanging raw data. In coming years, federated learning will be a fundamental approach to preserving the privacy of sensitive information and to reducing the need for data transmission to centralized cloud servers.

Federated learning enables edge devices to learn from a diverse range of data sources, which leads to more robust and accurate 'global' models. Most importantly, federated learning allows for continuous model updates and improvements, as devices can share new learnings and insights with their federated network in real-time.

However, federated learning is not the sole decentralized learning approach to developing Edge AI systems. There is also swarm learning, which takes inspiration from the collective behavior of social organisms (e.g., birds, insects) to create decentralized and self-organizing AI systems. The deployment of Edge AI networks based on swarm learning involves edge devices that work collaboratively to solve complex problems and make decisions by sharing information and insights in a decentralized manner. Contrary to federated learning, swarm-learning nodes (i.e., edge devices) share information completely decentralized without aggregating it in a centralized cloud. This approach enables Edge AI systems to dynamically adapt and evolve independently while at the same time improving their performance and efficiency in real-time.

In the future, such decentralized learning approaches will pave the way for more scalable, efficient, and privacy-preserving Edge AI solutions.



Capabilities comparison of cloud, on-device, and edge intelligence. Image credit: Viso.ai.

Heterogeneity and Scale-up Challenges

The future of Edge AI systems will include more scalable and heterogeneous systems. Specifically, the proliferation of the number and types of Edge AI systems will give rise to heterogeneous environments where different edge AI systems (e.g., TinyML, federated learning, embedded machine learning) will co-exist and interact in various deployment configurations. This will create additional challenges due to the need for integrating and orchestrating future Edge AI systems in complex AI workflows involving both cloud-based and cloud/edge components.

In this direction, workflow management and AI orchestration middleware for cloud/edge environments must evolve to address the peculiarities of edge AI systems. Future orchestrators must consider and balance trade-offs associated with the deployment and operation of Edge AI systems, including trade-offs concerning performance, latency, data protection, scalability, and energy efficiency.

The heterogeneity of future AI systems will raise the development complexity of end-to-end Edge AI systems. To alleviate this complexity, Edge AI vendors will likely offer more sophisticated tools for developing and deploying machine learning pipelines in cloud/edge environments. For example, they will offer novel AutoML (Automatic Machine Learning) environments that lower the time and effort needed to train, integrate, and deploy end-to-end applications.

Key Stakeholders and Their Roles

The future of Edge AI will be developed in the scope of an evolving ecosystem of stakeholders, which will interact closely to ensure that different developments fit together and reinforce each other. The heart of this ecosystem will comprise vendors of microsystems and edge devices and integrators of machine learning solutions for Edge AI systems. The traditional machine learning community will actively engage in the Edge AI ecosystem by advancing popular environments and tools to support future edge AI developments.

Research and development organizations will also have a prominent role in the Edge AI ecosystem, as they will drive the development of the technological innovations that will revolutionize AI. Moreover, the open-source community is expected to have an instrumental and pioneering role in prototyping, testing, and standardizing future Edge AI middleware and tools.

"Edge AI is like building a superhighway; as this highway is built, the possibilities that it unlocks are tremendous, just like how the highway system spawned economies," said Yanamadala. "The purpose of Edge AI is to unlock the potential of machine learning and AI applications. The end goal is to facilitate unleashing the power of ML combined with data. It might happen in stages, but that's the happy path."

Conclusion

As we conclude this report on Edge AI, we stand at the precipice of a transformative era in Artificial Intelligence. The rise of Edge AI technology presents us with unparalleled opportunities to shape the future of intelligent devices and systems. With its ability to process data locally, reduce latency, enhance privacy, and enable real-time decision-making, Edge AI is poised to revolutionize various industries and propel us toward a more connected, efficient, and intelligent world.

The importance of Edge AI in the current and future technology landscape cannot be overstated. Its impact is felt across domains such as autonomous vehicles, healthcare, industrial automation, and IoT deployments. By bringing AI capabilities directly to the edge devices, Edge AI empowers devices to operate autonomously, adapt to their surroundings, and make informed decisions in real-time.

As technology continues to advance, the growth of edge computing infrastructure and the development of specialized AI hardware accelerators will unlock even greater potential for Edge AI applications. We can anticipate the deployment of more sophisticated AI models on edge devices, enabling complex tasks such as natural language processing, computer vision, and deep learning directly at the edge.

Furthermore, the proliferation of 5G networks will complement Edge AI by providing high-speed, low-latency connectivity, further enhancing the capabilities of edge devices. This convergence of Edge AI and 5G will facilitate the seamless integration of intelligent devices and systems into our daily lives, enabling a world where autonomous vehicles navigate with precision, smart cities optimize resources, and healthcare systems deliver personalized care. The journey of Edge AI has only just begun, and the path ahead is illuminated with the promise of transformative advancements.

Acknowledgments

Many thanks to all our sponsors, writers, and supporters for their valuable contributions to the report.

Brian Fuller, Prathyusha Venkata, and all the people at Arm

Sean Hollister and the team at Sparkfun

Merlijn Linschooten and everybody at Axelera AI

Venkat Kodavati, Shay Kamin Braun, and the team at Synaptics

Jon Gallegos and Nandan Nayampally from BrainChip

The people at ST, especially, Louis Gobin, Mark Hopkins, and Pamela McCracken

Henrick Flodell and Alexandra Kazerounian From Alif Semiconductor

Martin Croome from Greenwaves Technologies

Stefano Implicito from Arduino

The writers John Soldatos, Miroslav Milovanovic, and Lydia Husser

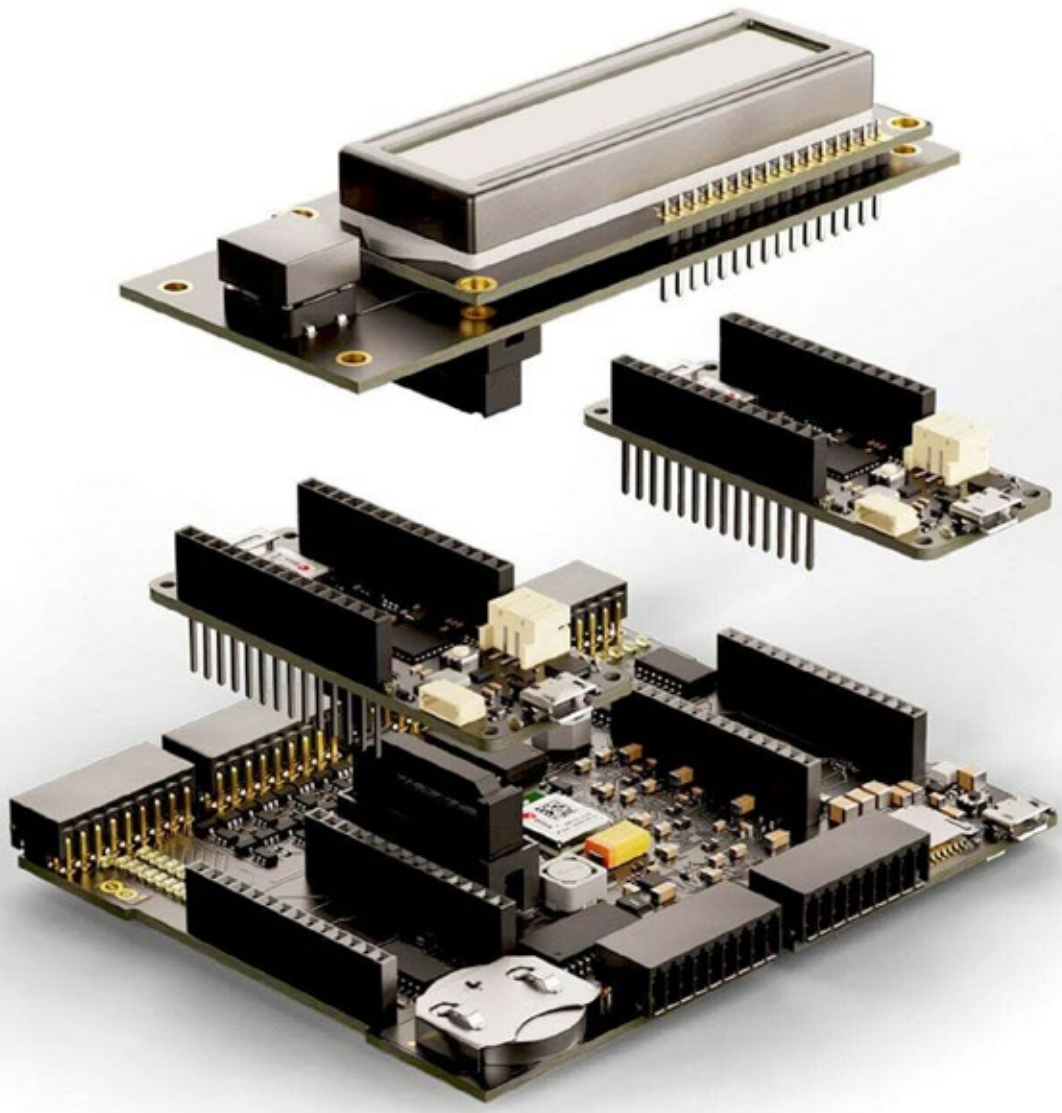
The Wevolver team, including Jessica Miley, Tasos Polygenis, Bram Geenen, and Richard Hulskes, and everybody who has helped with proofreading, knowledge sharing, and design.

The cover image of the report was created using generative AI tools. Image credit to Muthali Ganesh.

About the report sponsors

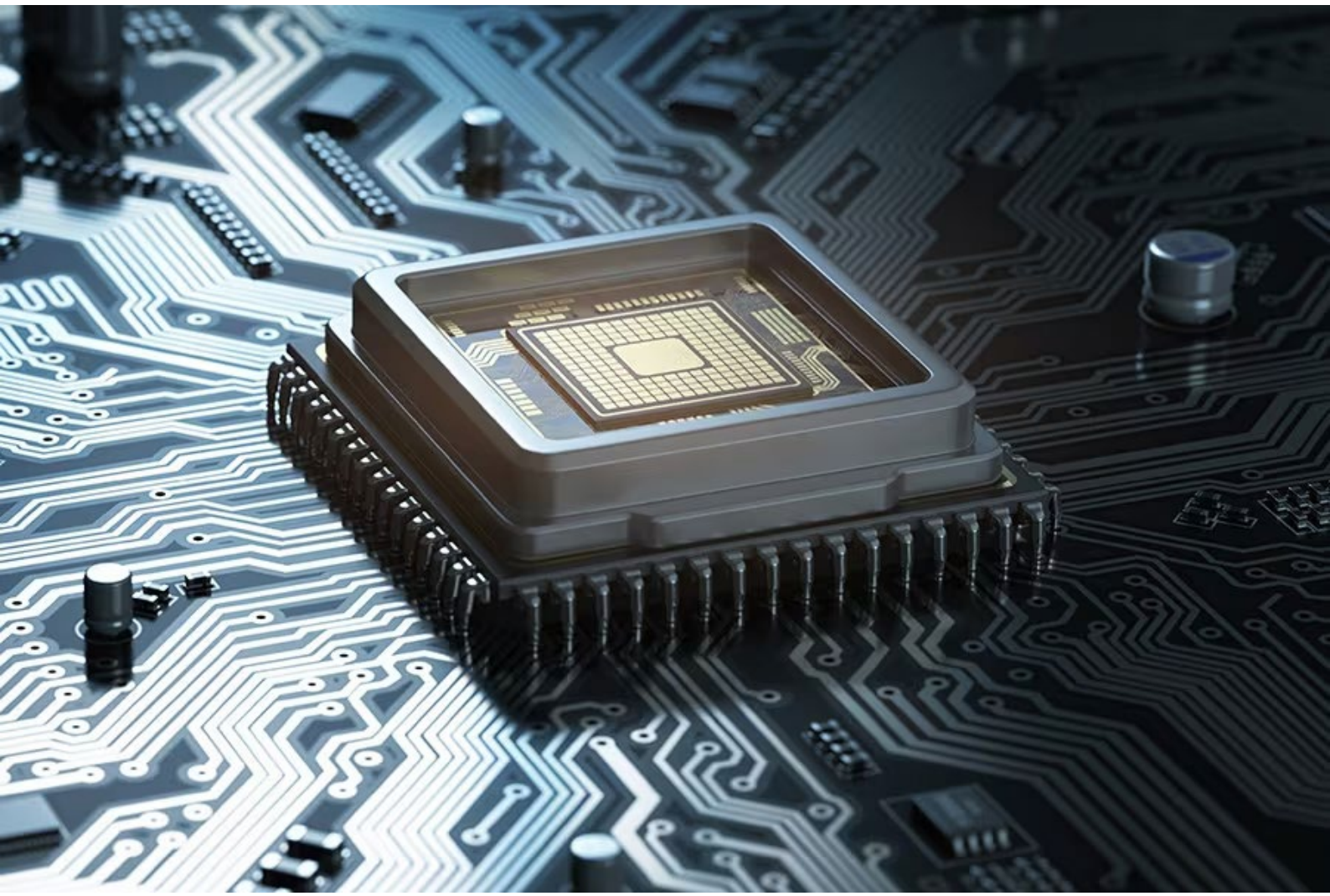
Arduino

Arduino is the leading open-source hardware and software company in the world, with a community of over 30 million active users. Born to provide an easy-to-use platform for anyone doing interactive projects, Arduino has reached a growing community and adapted to new needs and challenges, branching out into products for IoT, wearables, 3D printing, and embedded environments. Arduino Pro is the original all-in-one IoT platform. Companies can connect business logic with IoT sensor data in minutes using Arduino IoT Cloud and production-ready certified hardware. Built on Arm technology, the latest generation of Arduino solutions brings users the best of both worlds in terms of simplicity of integration and a scalable, secure, professionally supported service. Low-code IoT cloud application development allows you to iteratively refine your application and quickly assess the viability of wider IoT deployments without the need for expensive consultants or integration projects.



Arm

Arm technology is defining the future of computing. Our energy-efficient processor designs and software platforms have enabled advanced computing in more than 250 billion chips, and our technologies securely power products from the sensor to the smartphone and the supercomputer. Together with 1,000+ technology partners, we are enabling artificial intelligence to work everywhere, and in cybersecurity, we are delivering the foundation for trust in the digital world – from chip to cloud. The future is being built on Arm.



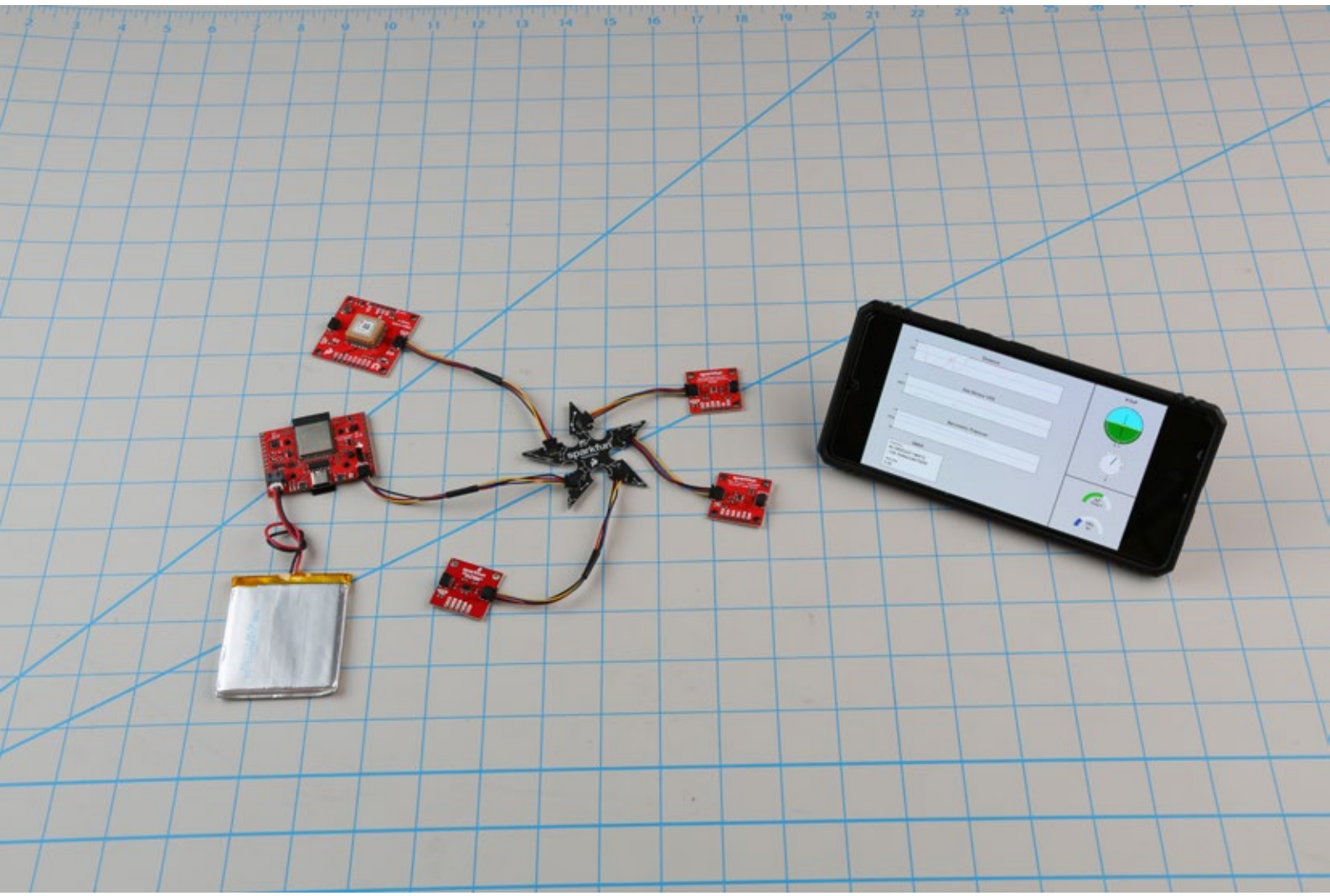
Edge Impulse

Edge Impulse offers the latest machine learning tooling, enabling all enterprises to build smarter edge products. Our technology empowers developers to bring more AI products to market faster and helps enterprise teams rapidly develop industry-specific solutions in weeks instead of years. Edge Impulse provides powerful automation and low-code capabilities to make it easier to build valuable datasets and develop advanced AI with streaming data. With over 75,000 developers and partnerships with the top silicon vendors, Edge Impulse delivers a seamless integration experience to validate and deploy with confidence across the largest hardware ecosystem. To learn more, visit edgeimpulse.com.


Heart rate
64 BPM 3 MINUTES AGO
173
54
12AM
6AM
12PM
6PM
5:42

Sparkfun

In 2003, CU student Nathan Seidle blew a power supply in his dorm room and found he could only order replacement parts in bulk. Since then, he has been on a mission to make electronics and supporting documentation available to individuals everywhere. Since the beginning, SparkFun has been committed to sustainably helping our world explore the edges of technology and achieve electronics literacy from our headquarters in Boulder, Colorado. No matter your vision, SparkFun's products and resources are designed to make the world of electronics more accessible. In addition to more than 2,000 open-source components and widgets, SparkFun offers online tutorials and training designed to help people use technology to solve problems, create products, and get creative. We're here to help you start something!



Alif Semiconductor

Alif Semiconductor is an industry-leading supplier of the next-generation Ensemble family of microcontrollers and fusion processors. The Ensemble family scales from single-core MCUs to a new class of multi-core devices, fusion processors, that blend up to two Cortex-M55 MCU cores, up to two Cortex-A32 microprocessor cores capable of running high-level operating systems, and up to two Ethos-U55 microNPUs for AI/ML acceleration. With multi-layered security, scalable performance, and next-level integration, Alif MCUs and MPUs provide a wide range of functionalities on a single monolithic die. The Alif Ensemble family delivers increased AI/ML efficiency, lowers power consumption, and keeps data safe, all while offering a scalable processor continuum.

Axelera AI

Axelera AI is providing the world's most powerful and advanced solutions for AI at the edge. Its game-changing Metis™ AI platform – a holistic hardware and software solution for AI interference at the edge – enables computer vision applications to become more accessible, powerful, and user-friendly than ever before. Headquartered in the AI Innovation Center of the High Tech Campus in Eindhoven, The Netherlands, Axelera AI has R&D offices in Belgium, Switzerland, Italy, and the UK, with more than 125 employees in 15 countries. Its team of experts in AI software and hardware hail from top AI firms and Fortune 500 companies.

BrainChip

BrainChip is a leader in edge AI on-chip processing and learning. The company's first-to-market, convolutional, neuromorphic processor, Akida™, mimics the event-based processing method of the human brain in digital technology to classify sensor data at the point of acquisition, processing data with unparalleled energy-efficiency and independent of the CPU or MCU with high precision. On-device learning that is local to the chip without the need to access the cloud dramatically reduces latency while improving privacy and data security. In enabling effective edge computing to be universally deployable across real-world applications, such as connected cars, consumer electronics, and industrial IoT, BrainChip is proving that on-chip AI is the future for customers' products and the planet.

brainchip
Essential AI

GreenWaves Technologies

GreenWaves is a fabless semiconductor company founded in 2014 and based in Grenoble, France. We design and market ultra-low power processors for energy-constrained products such as hearables, wearables, IoT & medical monitoring products. GreenWaves' system-on-chips enable companies to develop and bring to market products with new-to-world features enabled by state-of-the-art machine learning and digital signal processing techniques. Our leading-edge development tools enable audio and machine learning developers to harness the power of GAP processors productively. GreenWaves GAP9 processor powers features such as neural network-based noise removal and adaptive noise cancellation, multi-channel spatial sound, and listening enhancement technologies in next-generation earbuds and headphones with market-leading energy efficiency.

GREENWAVES TECHNOLOGIES

ST

ST is a global semiconductor leader delivering intelligent and energy-efficient products and solutions that power the electronics at the heart of everyday life. ST's products are found everywhere today, and together with our customers, we are enabling smarter driving and smarter factories, cities, and homes, along with the next generation of mobile and Internet of Things devices. By getting more from technology to get more from life, ST stands for life.augmented.

Synaptics

Synaptics (Nasdaq: SYNA) is changing the way we engage with connected devices and data, engineering exceptional experiences throughout the home, at work, in the car, and on the go. Synaptics is the partner of choice for the world's most innovative intelligent system providers who are integrating multiple experiential technologies into platforms that make our digital lives more productive, insightful, secure, and enjoyable. These customers apply Synaptics' differentiated, AI-enhanced technologies for advanced connectivity, video, vision, audio, speech, touch, display, biometrics, and security processing.

About the report partner

tinyML Foundation

The tinyML Foundation is a non-profit professional organization focused on supporting and nurturing the fast-growing branch of ultra-low power machine learning technologies and approaches dealing with machine intelligence at the very edge of the cloud.

About the writers

Samir Jaber

Samir Jaber is an SEO & Content Specialist with a background in engineering, nanotechnology, and scientific research. Samir has comprehensive experience working with major engineering and technology companies as a writer, editor, and digital marketing consultant. He is an expert in content management and strategy, particularly in the engineering and tech fields. He is a featured author in 30+ industrial magazines with a focus on IoT, nanotechnology, materials science, engineering, and sustainability. Samir is also an award-winning engineering researcher in the fields of nanofabrication and microfluidics.

John Soldatos

John Soldatos holds a Ph.D. in Electrical & Computer Engineering from the National Technical University of Athens (2000) and is currently an Honorary Research Fellow at the University of Glasgow, UK (2014-present). He was Associate Professor and Head of the Internet of Things (IoT) Group at the Athens Information Technology (AIT), Greece (2006–2019), and Adjunct Professor at the Carnegie Mellon University, Pittsburgh, PA (2007–2010). He has significant experience working closely with large multi-national industries (e.g., IBM, INTRACOM, INTRASOFT International) as an R&D consultant and delivery specialist while being a scientific advisor to various high-tech startup enterprises. Dr. Soldatos is an expert in Internet-of-Things (IoT) and Artificial Intelligence (AI) technologies and applications, including IoT/AI applications in smart cities, finance (Finance 4.0), and industry (Industry 4.0).

Miroslav Milovanovic

Miroslav Milovanovic has a Ph.D. in Computer Science and Electrotechnics and works as an assistant professor at the Faculty of Electronic Engineering. He teaches courses strongly connected with Data Science, the Industrial Internet of Things, Modern Control of Industrial Processes, and Intelligent control. Additionally, Dr. Milovanovic is the Chief of the Laboratory for Intelligent Control in the Control Systems Department. His instructional expertise encompasses courses strongly connected with Data Science, the Industrial Internet of Things, Deep Learning, Machine Learning in Python, Modern Control of Industrial Processes, and Intelligent Control. By imparting practical knowledge in these areas, he empowers students to excel in their chosen fields. With over 45 published scientific papers, his research primarily revolves around Data Science, Deep Learning, and their practical applications.

Lydia Husser

Lydia Husser is a technical writer in robotics and a freelance writer captivated by innovative science and engineering. She holds a degree in Computer Science and credentials in electronics technology. Her background includes work in robotics, communications hardware, and developer documentation.

About the designers

Jelena Krco

Jelena is an architect from Bosnia and Herzegovina. She earned a Master's in Architecture and Urban Design from the Faculty of Technical Sciences in Serbia in 2017. While working on her thesis, she developed an interest in art and architecture in post-conflict societies and has been involved in many research projects and workshops on this topic throughout the Balkan region. She is an experienced illustrator who has been working as a freelance professional in the field for several years. Currently based in Turin, Italy, Jelena is also pursuing her second Master's degree in Sustainability and Climate Change Mitigation at Politecnico di Torino.

Eszter Tóth

Eszter is a textile and graphic designer from Budapest, Hungary. She graduated from Moholy-Nagy University of Art and Design. Afterward she was building her own sustainable accessory design brand called Müskinn for several years. She also obtained a degree in art therapy. She has been working as a freelance graphic designer in such projects as independent movies, advocacy campaigns for NGOs, and making illustrations for pedagogy projects. She is currently working as an art therapist and a freelance graphic designer.

About Wevolver

Wevolver is a digital media platform & community dedicated to helping people develop better technology. At Wevolver we aim to empower people to create and innovate by providing access to engineering knowledge.

Therefore, we bring a global audience of engineers informative and inspiring content, such as articles, videos, podcasts, and reports, about state of the art technologies.

We believe that humans need innovation to survive and thrive. Developing relevant technologies and creating the best possible solutions require an understanding of the current cutting edge. There is no need to reinvent the wheel.

We aim to provide access to all knowledge about technologies that can help individuals and teams develop meaningful products. This information can come from many places and different kinds of organizations: We publish content from our own editorial staff, our partners like MIT, or contributors from our engineering community. Companies can sponsor content on the platform.

Our content reaches millions of engineers every month. For this work Wevolver has won the SXSW Innovation Award, the Accenture Innovation Award, and the Top Most Innovative Web Platforms by Fast Company.

Wevolver is how today's engineers stay cutting edge.

Address

Plantage Middenlaan 62
1018 DH Amsterdam
The Netherlands

Contact

@wevolverapp
www.wevolver.com
info@wevolver.com