Simplifying AI Deployment from the Cloud to Edge and Endpoint

Suad Jusuf, Director, Product Marketing and Strategy - AI COE, Renesas Electronics Corporation

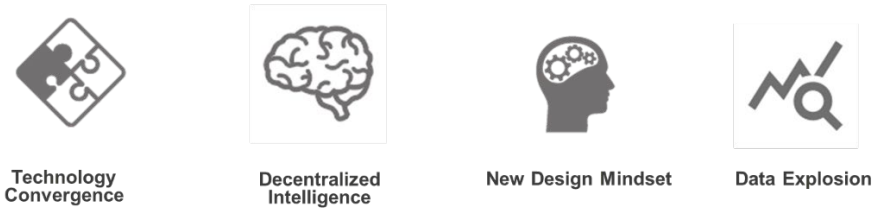
Abstract

Artificial Intelligence (AI) is transforming every aspect of life. It is enhancing quality in industrial applications, enabling smart home systems, monitoring our safety as we work and play. Advances in technology have allowed us to run complex machine learning algorithms to tackle unique problems allowing those to be implemented also on embedded devices used in our daily life in home and industry. To enable scalable intelligence at all levels of the network, a decentralized intelligence architecture is needed. This means running cloud-independent inference engines on power-efficient or tiny computers within the edge and endpoints.

In this white paper, we will address the trends driving a decentralized intelligence model, the key application categories benefitting from AI, and how Renesas' approach simplifies design and deployment for AI / ML developers to help surmount commonly encountered challenges.

Overview

Today there isn't any doubt that AI is changing the aspects of our everyday life both on and offline mostly in the background. The AIoT or Artificial Intelligence of Things is recognized as a transformational megatrend driven by four underlying market dynamics which include technology convergence, decentralized intelligence, a new design mindset, and an explosion of data.



- Technology convergence – IoT, AI, 5G maturing at roughly the same time
- Decentralized intelligence – Tremendous benefits of a distributed intelligence model
- New design mindset – AI disrupting system design approaches
- Data explosion – Endpoint data creation expected to grow 85% from 2017-2025*

*Source: IDC Research

Why Decentralize Intelligence?

Up to this point, the Internet of Things (IoT) has been built on a cloud-centric intelligence architecture. Based on overall experience and analysis over the past several years, we've learned the lesson we had to learn. Namely that widely distributed applications generate a vast amount of data. It has been clearly identified that this enormous amount of available data is unable to be handled, analyzed, and used. Furthermore, it has increased energy consumption and overall system costs. Even worse, the intelligence of AI was not intelligent enough to act/react when the cloud connection to the application was missing or interrupted. We don't need to mention the inability of any real-time response in this case. So, what we have realized is simply that not every application needs to be or should be served with cloud-centric intelligence. It doesn't mean that the data exchange to the cloud must be abandoned, instead it means we need to remove this unnecessary dependency and establish a more suitable decentralized approach that can enable the use of intelligence right where it's needed. Decentralizing intelligence, especially in the context of AI, refers to distributing computing power and decision-making capabilities across a network of devices rather than relying solely on centralized systems and cloud architecture.

Key benefits of a decentralized intelligence:

Supports Scalability

Allows the addition of resources to the network as demand grows and enables the system to handle increasing workloads without reaching critical bottlenecks or performance degradation.

Enables Real-time Response

Latency can be drastically reduced by distributing intelligence closer to where data is generated or consumed, enabling real-time processing and optimization of time-critical applications.

Enhances Security and Data Privacy

Enhances privacy and security by minimizing the need to transmit sensitive data to centralized servers. Instead, data can be processed locally on edge devices tightly coupled to hardware root-of-trust, reducing the risk of unauthorized access or data breaches.

Reduces Costs and Increases Network Agility

Enables edge computing, where data is processed and analyzed closer to the source—such as IoT devices or sensors—allowing for faster insights and actions along with reduced bandwidth usage and costs associated with transferring large amounts of data to centralized servers.

Merging AI and IoT

The Artificial Intelligence of Things represents a powerful synergy transforming industries and enabling innovative applications across various market segments and domains. IoT systems can be thought of as an interplay of one or more core technologies as shown in *Figure 1* below.

The vast amount of data generated from sensors, machines and connected devices can now be analyzed in real-time to extract valuable insights, detect patterns, and make predictions targeting process optimizations, equipment failure detection and proactive maintenance planning. All of which are exactly the goals of a decentralized approach which will allow us to embed AI directly into IoT or edge computing devices so systems can adapt to changing conditions, autonomously make decisions and respond to events in real-time.

Personalized experience, tailored recommendations, notifications, and services offered to users, create enhanced and more engaging and relevant experiences. Overall, AI and machine learning (ML) can be viewed as enablers bridging underlying technologies and applications, offering tremendous potential to improve efficiency, enhance user experience and drive innovation. AI / ML use can be classified broadly into three main categories – voice which deals with spoken voice and linguistics, vision which deals with camera, machine vision and robot vision and, finally, real-time analytics mainly addressing time-series data that is captured from sensors or sensor-less using the available system parameters.

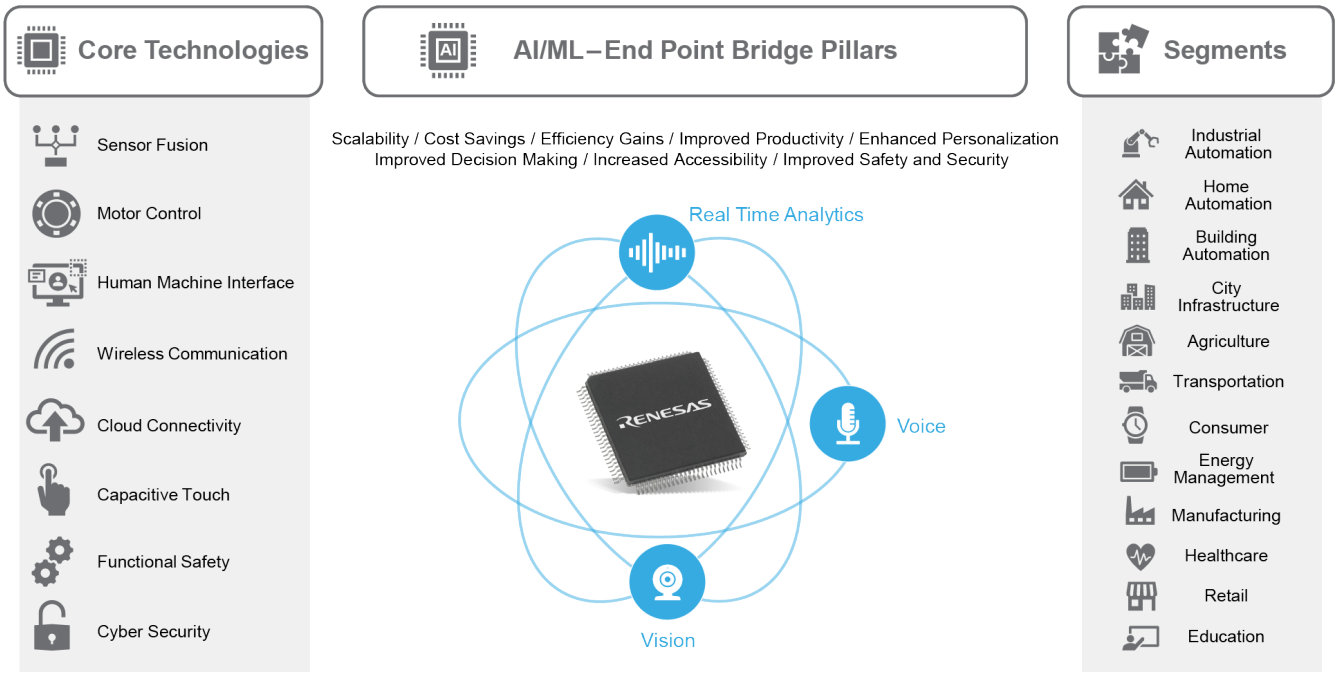


Figure 1 Core IoT technologies that can be enabled by AI and machine learning

Renesas offers a comprehensive solution stack for AI / ML developers, as well as a broad product portfolio spanning sensing, connectivity, computing, and actuation to cover all layers of the IoT. In addition, software tools, solutions, and a broad partner ecosystem are purpose-built to accelerate your AIoT designs.

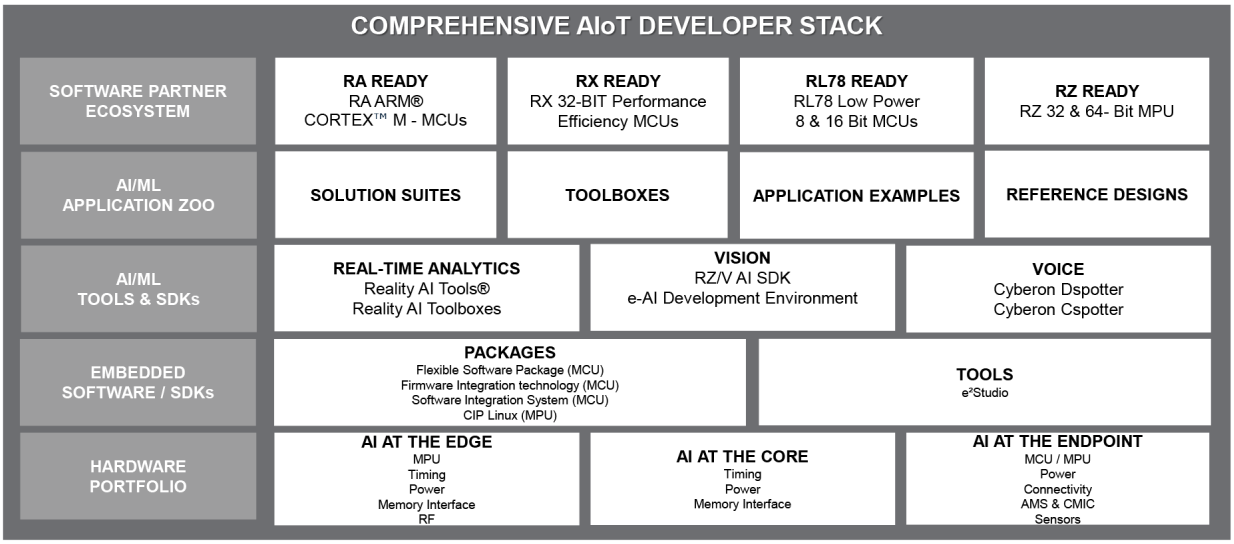


Figure 2 Renesas multi-layer developer stack

From the multi-modal needs of applications and overall complexity, users often have the challenge of selecting the best fitting and cost-optimized device for their system. Here Renesas offers a scalable lineup within the MCU/MPU product families, helping to select the right one for a dedicated AI / ML application while also fulfilling the requirements from a performance vs multimodality and complexity perspective as depicted in Figure 3.

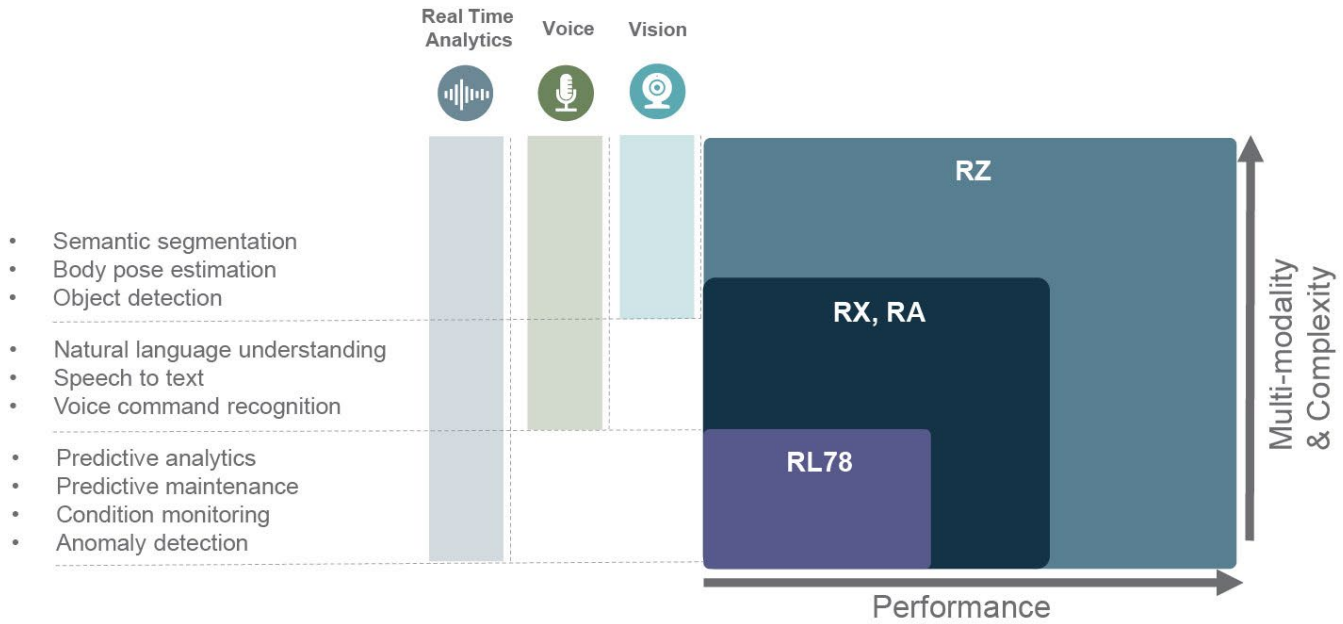


Figure 3 Renesas MCU / MPU portfolio addressing AI / ML specifications

We've addressed the Renesas AI / ML view and segmentation into three main pillars, as well as a bit about the overall stack offering and the scalable device lineup. Now, let's look at the Renesas coverage for the three main AI / ML pillars and see the available resources and how they simplify AI deployment.

AI / ML Pillars

Voice

As speaking is a natural and intuitive way for humans to communicate, a Voice User Interface (VUI) allows us to interact with devices using our natural language. It reduces the learning curve and makes the interaction more user-friendly. VUI allows us to control various smart devices, such as thermostats, lights, appliances, making the overall experience more seamless and convenient. Despite significant advancements, VUI development challenges related to accuracy, robustness, and privacy still need to be considered. Otherwise, the user-friendly aspect can easily disappear in frustration. Support of different languages and accents is a requirement for global systems development. And that's where the journey begins with the extensive preparation process and data collection for each language. Training, evaluation, testing and optimizing can become a prolonged process with preprogrammed headaches.

And we cannot forget to ensure critical user privacy and data security, given the sensitive nature of spoken interactions. Of course, optimal designs will strive for a cost-effective, decentralized endpoint on an embedded device.

The complete Renesas solution offers accelerated integration of your own VUI in your application. The development environment has already tackled the burdens and represents an easy-to-use embedded hardware platform for Voice User Interface solutions without extensive coding experience or in-house expertise. It is based on small, general purpose MCUs and even offers a RISC-V ASSP option as a standalone solution.

From a software perspective, the DSpotter Modeling Tool from Cyberon is provided as a GUI development environment. It enables simple, flawless development of an endpoint voice command recognition (VCR) or Natural Language Understanding (NLU) user interface. Pretrained models for 44+ languages eliminate the burden of extensive data collection and training the model accordingly. Customized commands can be easily generated by text input and the user can have offline performance tuning and testing available.

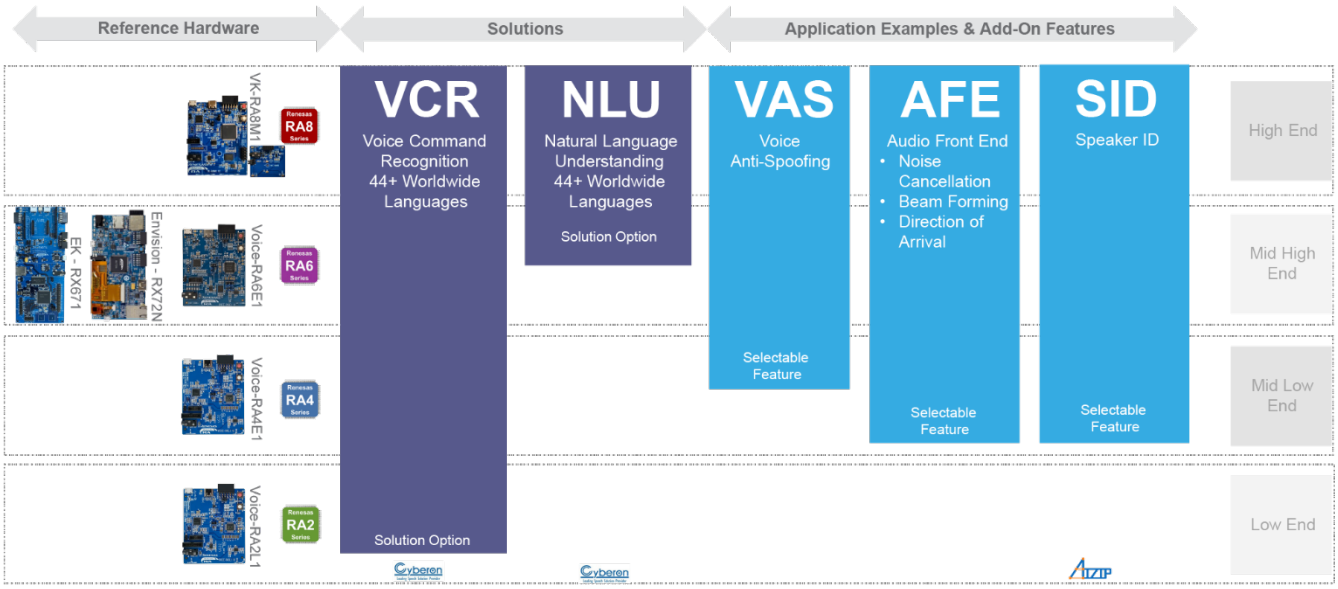


Figure 4 AI / ML Voice solutions

Additionally, Renesas provides enhanced features and application examples like voice anti-spoofing and audio front end, as well as Speaker Identification from our Partner Aizip, as depicted in Figure 3. These feature sets can be added to the VCR or NLU to enhance the capabilities of the VUI to address safety, noise resistance and personalization along with ensuring user privacy and data security.

From the value proposition standpoint, here are some important take aways:

- Extensional application examples as add on features
- Voice anti-spoofing enables safety by distinguishing between real human and recorded voice
- Audio Front End enhances the audio / noise resistance capabilities
- Noise cancellation and beamforming included
- Speaker ID enables safety and personalization with user identification

Furthermore, material including documentation, webinars, videos, even hand-on session training is available on our web site to support you and your new VUI development. For more information, please visit the Renesas Voice Solutions page.

Vision

Vision is enabled by camera-based systems and provides machines with the ability to interpret and understand visual information from images and videos. As a subset of AI, computer vision focuses on algorithms and techniques that enable machines to analyze, process, and extract meaningful insights from visual data.

Enabling a wide range of technological innovation, vision systems provide diverse approaches, such as image segmentation, object detection, facial recognition, edge detection, pattern detection, image

classification, and feature matching. It is widely used in the manufacturing industry, healthcare, transportation, city & infrastructure, building automation and everyday life.

Key challenges are related to scalability and efficiency, data availability and ensuring quality and robustness, just to name a few. The crucial step here is addressing the question of how to tackle those from the very beginning in order to reduce burdens like the duration of evaluation and preparation of the substantial amounts of data needed for training. As if that were not enough, implementation on an embedded device must now be considered. Typical concerns for resource limitations, performance and power consumption still need to be addressed. These can all be managed, including initial cost reduction, shortening the transition from development to deployment and accelerating time to revenue. So where do we begin and how?

Getting started

Let's begin with the selection of the hardware. For AI / ML systems, as with others, designers need hardware platforms offering scalability for factors such as processing power, memory, connectivity options, size, and power consumption. Evaluating off-the-shelf embedded systems, development boards or custom hardware solutions based on project requirements is more than given and important to address the target of our mission.

Renesas provides comprehensive offerings that includes a wide range of scalable MCU and MPU product families. An environment of reference hardware, application examples, pre-trained models, SDKs and tools such as the AI Navigator help guide developers to accelerate the evaluation and development of Vision-based AI applications.

Widely available devices feature one or more camera interfaces, dedicated graphical units (2D / 3D), high-performance image signal processor (ISP) supporting 4K/30fps and dynamic reconfigurable processor units known as DRP and DRP-AI (DRP-AI3) as AI Accelerators, in addition to advanced power management systems. Figure 3 shows the device series arranged by vision approach, complexity and multi-modality.

The RA8 series of MCUs is equipped with Arm® Cortex®-M85 (Armv8.1-M architecture) core enhanced with the Helium™ vector processor which significantly uplifts the performance for ML and signal processing (DSP) applications. It delivers over 3000 CoreMark points at 480MHz operation. The RZ family of MPUs and the RZ/V2H device catapults users to the high-end of computing and computer vision. It comes with quad Cortex-A55 at 1.8GHz, dual Cortex-R8 real-time processor at 800MHz and single Cortex-M33 at 200MHz plus DRP-AI3 and DRP unit. With support for spars, pruning modeling techniques, it also features low-power consumption with best-in-class thermal performance at amazing 10TOPS/W performance/power consumption.

For more information about key devices and specific AI accelerators and enhanced IP, visit the RA8D1 page, RZ/V series page or Renesas DRP-AI page.

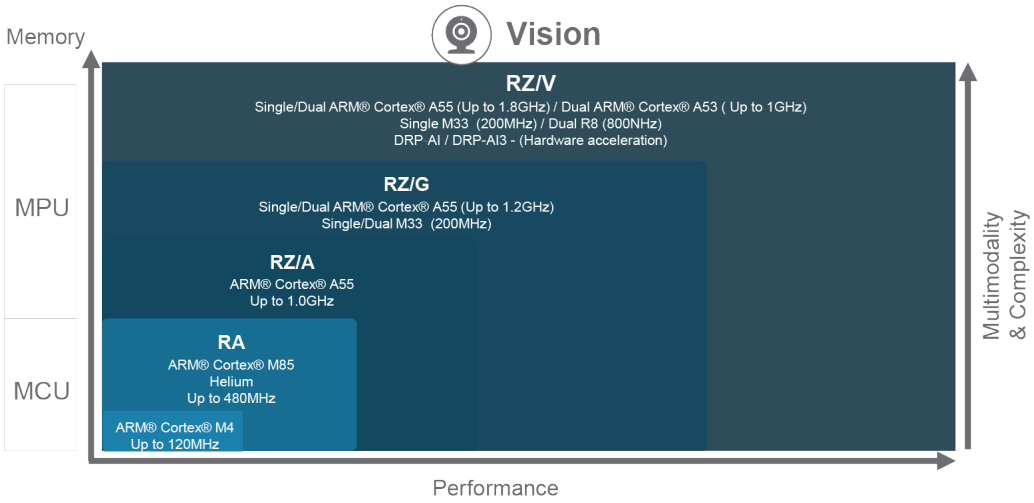


Figure 5 MCU / MPU options to address performance and complexity of vision systems

On the microcontroller side, the RA8D1 evaluation kit comes with the Flexible Software Package (FSP) and the embedded development environment e²studio. Additional vision-based application examples are available from our Ecosystem partners for different use cases, such as driver condition recognition, camera-based QR code scanning, person access system with vision based anti-spoofing, just to name few. Visit the RA AI partner page for more information and a complete list of all the available application examples, including people and object recognition, detection and access solutions.

On the microprocessor side, Renesas offers a development environment to support different levels of expertise in AI vision design with a combination of comprehensive tools and a wide range of application examples with pre-trained AI models.

The RZ/V2H evaluation kit supports standard software packages and enables implementation of low power consumption AI inferences and video streaming. Additionally, it supports optional functions like the image signal processor (ISP), 3D graphics engine (GE3D), and Trusted Secure IP. It comes with the RZ/V2H AI Software Development Kit (AI SDK) as an AI application development environment. The Renesas RZ/V2H-EVK provides more information and resources including documentation, video tutorial, and others.

DRP-AI TVM (powered by the Edgecortix MERA™ compiler framework) is a tool that generates runtime executables for AI from trained AI models for RZ/V series devices. It enables a specific separation of the executables to run on the CPU and the DRP-AI3 for highly optimized inference execution.

DRP-AI Translator enables the translation of the AI model into an DRP-AI library which completely runs on the DRP-AI. The SDK package includes a complete image with a compiler and all the libraries needed for

Linux cross-compilation. It enables execution of an application compiled with e2studio on the evaluation board.

AI Navigator is a plug in for e²studio and includes a transfer learning tool to retrain new classes of classifications within the available AI models. It allows integration and operation of various functions needed to develop AI and provides:

- Selection of AI applications from the Renesas AI application zoo and download of corresponding e2studio projects
- Customization of AI models for supported AI applications with their own datasets with the transfer learning feature
- Conversion of AI models to executable files. The RZ/V tool allows conversion to DRP-AI executable code using TVM

Developers can choose one of three options to get started with Renesas tools based on their level of expertise and where they are in the development journey:

Immediate start: accelerate the process with available, pre-trained models within the application zoo, select from among 50+ different models to generate your executable file using the AI SDK and an RZ/V evaluation kit.

Bring your own Data (BYOD): for a more advanced approach where the available models need some customization based on available data set, using the transfer learning tool and retrain new classes of classifications. The DRP-AI TVM generates highly optimized runtime executable and the AI SDK generates the compilation for the RZ/V evaluation kit.

Bring your own Model (BYOM): if you have one you would like use to evaluate the RZ/V capabilities, this path enables you to immediately generate the runtime executable through the DRP-AI TVM and AI SDK.

The process of the integrated flow, as depicted in *Figure 6*, incorporates all the above tools and enables selection for the right path for your vision-based AI use case.

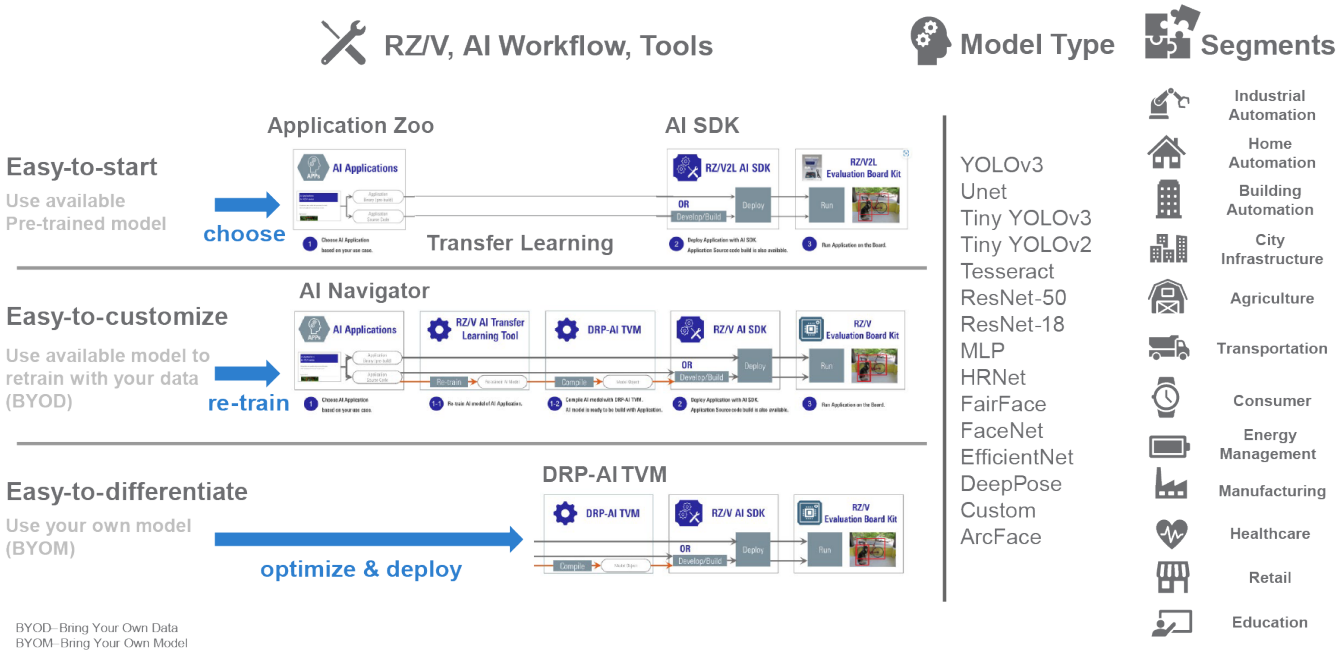


Figure 6 Selectable path for the RZ/V AI workflow, tools, and supported model types

From lighting control depending on room occupancy to security area instructions or hand gesture recognition and touchless elevator control, the application zoo includes a variety of different use cases as applications with pre-trained models that are ready for evaluation, testing and further work, depending on your own application requirements. For more information and to start working with Renesas AI Vision solutions, please visit the RZ/V AI GitHub site.

Real-time Analytics (RTA)

This data driven engineering discipline enables a process of preparation, proceedings, and analyzing time series data as soon as available. Time-series data are simple measurements or tracked conditions/events, monitored, sub-sampled, and accumulated over time from different sources and sensors. It allows AI / ML models to learn from temporal dependencies, identify patterns, inconsistencies, and track changes over time. This is essential for applications where past observations influence future outcomes, such as anomaly detection and sequential decision-making. RTA empowers the opportunity for feature engineering, which can be derived from temporal patterns and decision-making in applications where timely insights are critical and can adapt to evolving conditions and learn from new observations, ensuring robustness and accuracy in dynamic settings.

Renesas RTA solutions address non-visual time-series data and provide an extensive development environment, helping starters or longtime pros to accelerate embedded deployment and cut months, or

even years from an R&D cycle. Easy orientation within the full product lineup provides powerful tool environment, reference kits, solutions, toolboxes, and various application examples.

Reality AI Tools, the Renesas AI / ML flagship software, is a cloud-based edge and endpoint AI development environment. It combines advanced signal processing and machine learning which allow engineers to generate and build TinyML / edge AI models using powerful feature space ML path including SVM's and NN techniques. The tool supports all Renesas processor cores and delivers automatic parameters tune capabilities of the machine learning algorithm based on discovery features that leads to accuracy improvements and greater computational efficiency. Users can explore sensor data and generate optimized models for small MCUs automatically and generate the embedded code targeting the best fitting device out of the full range of device portfolio.

The tool contains very helpful analytics to find the best sensor, combination of sensors or the best locations for placement and automatically generates the component specs. Includes fully explainable model functions in terms of time/frequency domain. For further information, visit the Reality AI Tools page.

Furthermore, a seamless developer workflow has been established for rapid AI model prototyping with Renesas evaluation kits integrated with e²studio and Reality AI Tools. This integration enables users of e²studio to easy collect and upload data to Reality AI Tools, work in the environment to develop, train, and optimize the ML blocks and export the embedded code back into their project for live testing.

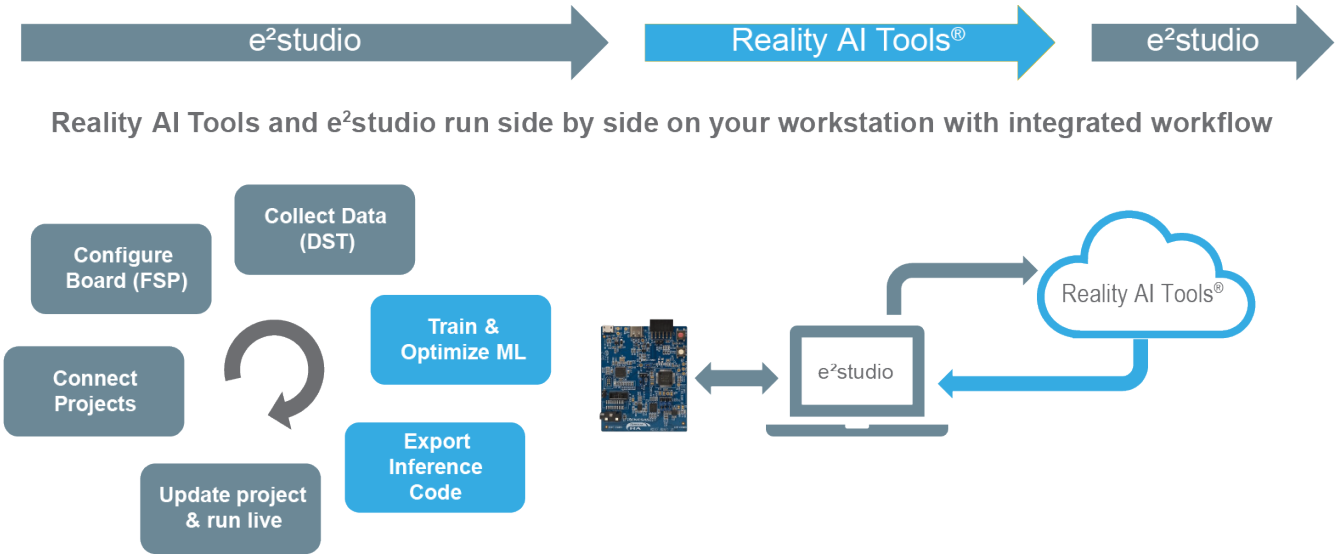


Figure 7 Integrated workflow between e²studio and Reality AI Tools

Alongside the Reality AI Tools suite, Renesas offers function-dedicated products accelerating the development to deployment phase within specific industrial application fields as depicted in Figure 7. These solution suites combined with dedicated Renesas support services assist developers from the very beginning to the end of the deployment.

- **Reality AI Tools** – Automatically explores sensor data and generates optimized models
- **Automotive Sound Recognition** – Safety Warning System (SWS) combines hardware and software to give passengers a new level of protection
- **RealityCheck™ HVAC** – Complete framework to enable smart, self-diagnosing HVAC systems
- **RealityCheck™ Motor** – Advanced software toolbox enables predictive maintenance and anomaly detection
- **RealityCheck™ AD** – Anomaly detection for monitoring factory and process-industry assets

Application examples and reference kits

Interested enthusiasts and developers who want to immediately evaluate and try out any of the RTA AI / ML use cases can leverage a suite of application examples and reference kits. From anomaly detection and presence detection to asset movement recognition, Renesas offers a variety of pre-built reference applications, enabling an easy and quick practical introduction to the subject of interest as depicted in *Figure 8*. A wide range of examples across different applications are available, in combination with the reference kits supported or cloud-only for testing and evaluation.

Cloud Only	Hardware	Reference Kit
Presence Detection	Voice Anti-Spoofing	Voice-RA6E1 / VK-RA8M1
Urban Scene Classification	Asset Movement Recognition	AIK-RA4E1 / AIK-RA6M3 / FPB-RL78 / CK-RA6M5
Pump Condition Monitoring	Friction Detection	MCK-RA6T2 / MCK-RA8T1 / RSSK-RX66T
Activity Detection Run vs Walking	Unbalanced Load Detection	MCK-RA6T2 / MCK-RA8T1 / RSSK-RX66T
Vibration Analysis on Rotating Shaft	Floor Type Detection	MCK-RA6T2
Fan Blower Condition Monitoring	Shaft Alignment and Unbalanced Load Detection	MCK-RA8T1

Figure 8 outlines the various available application examples within Real-Time Analytics AI / ML field

Reality AI Tools supports a variety of different evaluation and development kits with different devices for different applications.

The default kits package includes the full collateral material from schematics, BoM, documentation to software packages with dedicated BSP's, application notes and more.

Technology-driven boards like the Cloud Kit (CK), Motor Control Kit (MCK) and others, are intended to support the technology they are build for and provide dedicated capabilities demonstration of the device

itself within the technology requirements and serves as a proof of concept for further development as an application reference. Within the technology driven boards, Renesas has recently released scalable AI Kits (AIKs) to support AI / ML evaluation and development of vast of different use cases tackling complex tasks and challenges across various domains and also enables multimodal implementation. *Figure 9* shows the AIK-RA6M3 and the AIK-RA4E1 kits with communication interfaces to enable a full flexibility in selection of sensors, adaption to different systems and enabling centralized or decentralized node.

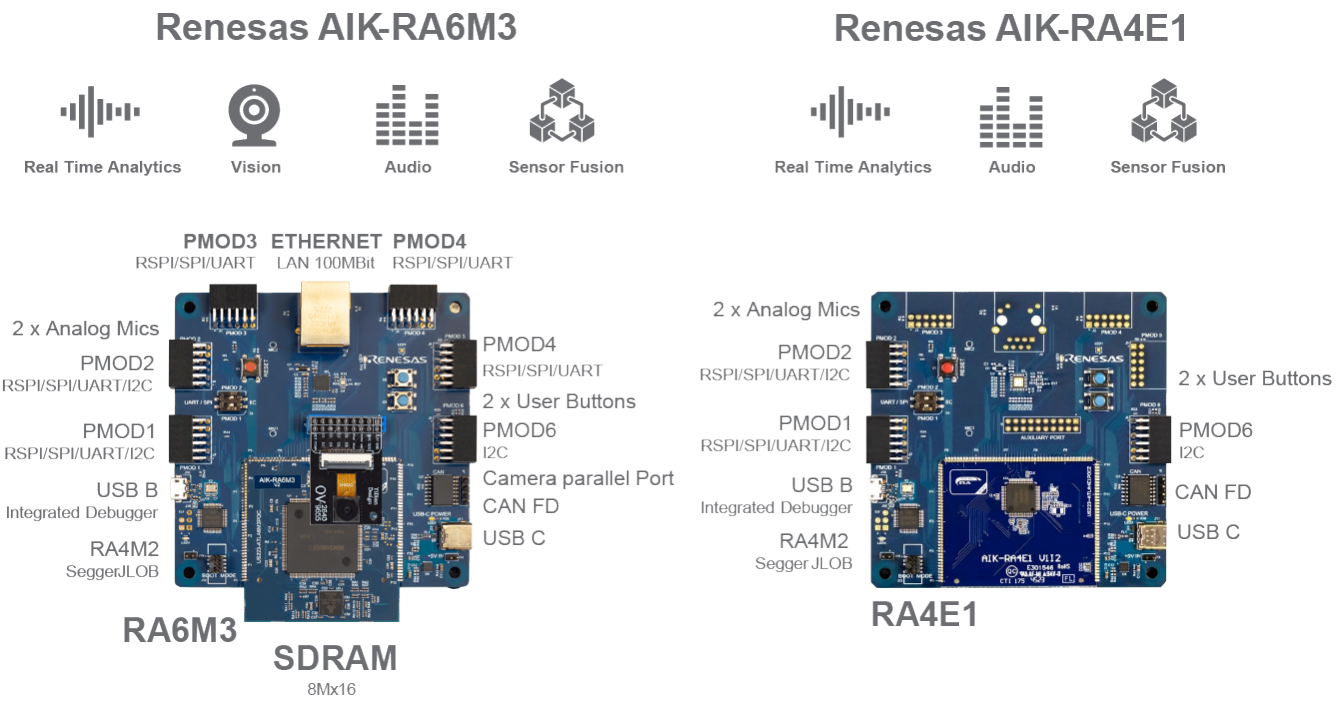


Figure 9 outlines the characteristics of the AI Kits

Each kit comes with highly tuned AI / ML models combined in a multi-modal solution with optimized CPU performance requiring very low memory resources. In combination with Reality AI Tools, users have a powerful AI / ML environment to kick off the development of RTA-based applications and build a flexible system. Users are equipped with camera, Pmod acceleration sensor module and Pmod display. An out-of-the-box example for asset movement recognition is included.

As an example of a multi-modal approach including vision, Renesas recommends the Person Access System (PAS) application example from our partner Aizip.

For more information on the AI Kits, visit the individual AIK-RA6M3 and AIK-RA4E1 pages.

Broad Partner Ecosystem

The Renesas Ready Partner Network is an extensive, curated network of trusted technology partners delivering commercial-grade building blocks that work out-of-box with Renesas products. These solutions are designed to help accelerate the development of IoT capabilities around core technologies such as security, safety, artificial intelligence/ machine learning, connectivity, cloud, sensing & control and human machine interface. Whether adding voice capability to an IoT product or applying machine learning to a specific design problem, these pre-developed solutions will significantly help reduce design complexity and speed time to revenue.

For more information, see the AI ecosystem solutions page.

Example of AI Deployment in Arc Fault Application

Let's take a look at a real-world example of AI / ML technology in an arc fault circuit interrupter (AFCI) for PV/ESS systems enabled and deployed with Reality AI Tools at a solar panel manufacturer.

Quick Overview: Arc fault detection is very important technology to maintain and guarantee the safety of power systems and is essential for manufacturing practical power systems for real-world applications.

Needs and Challenges: Arc-faults in PV systems may occur caused by various issues, such as faulty components, installation errors, mechanical damage or aging occurring after installation. Common examples causing arc-fault are damaged, pinched, or abraded conductors. Furthermore, loose or separated connections or terminations can lead also to arc faults. Additionally, switching noise from a grid-connected inverter makes detecting arc fault conditions even more difficult. There are various algorithms already available and implemented to solve arc fault detection, but the pain points remain, namely the maximum achievable accuracy of fault detection and the stability of detection under different conditions. Some system algorithms do have arc fault detection capabilities but typically at only 50% accuracy, which is clearly a disadvantage and poses a major safety threat.

Solution and Benefits: Renesas provided a comprehensive solution using AI / ML technology running on an RA6M4 MCU. With the data available in combination with additional data collected at the facility/lab, the Reality AI Tools deployment effectively generated the best fitting compact model with the lowest footprint, at ≤8K RAM, 23K Flash and reach accuracy of ≥99.8%. The flexible and scalable approach for this application enables further overall CPU performance and supports various system parameter-level scenarios up to maximum of 200A by maintaining the required effectiveness and stability.

Benefits:

- Performance optimized solution
- Compact model with low footprint
- Improved accuracy and stability
- Cost optimized and platform capable

Figure 10 shows an example of an AFCI implementation where the RA6M4 MCU combined with Reality AI Tools software enabled and integrated AI / ML approach to provide arc fault detection.

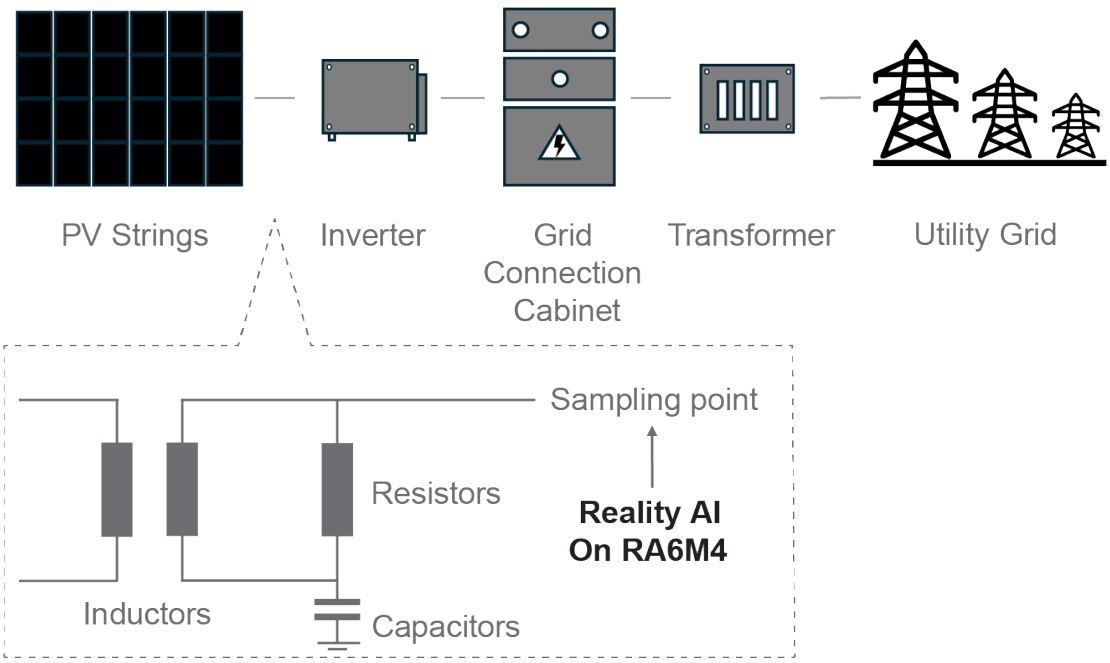


Figure 10 outlines a typical AFCI PV/EVSS system topology example

Renesas Approach

Renesas aims to accelerate your development and deployment of intelligent systems to enhance efficiency, safety, and user experience. Combining expertise in semiconductor design, embedded systems, and real-time control with AI processing capabilities, Renesas is actively investing in AI technology and solutions to address the growing demand for AI-driven applications across various industries, including industrial automation, automotive, consumer electronics, and IoT.

The Renesas approach integrates:

- Wide range of comprehensive hardware platforms with scalable MCU and MPU families
- Embedded tools and solution packages
- Comprehensive developer stack for AI / ML
- Tools and workflows suitable for multiple developer journeys – bring-your-own-model, transfer learning, bespoke consulting and more
- Rich library of easy-to-find solutions – application examples, toolboxes, solution suites, hardware reference kits
- Broad ecosystem of trusted partners offering commercial-grade building blocks

Conclusion

Renesas is broadly committed to providing innovative semiconductor solutions for a diverse range of applications and industries. With continued investment in technology development, product innovation and customer support, we will address evolving market needs and enable the next generation of intelligent systems.

Renesas edge AI solutions enable on-device AI processing for applications requiring low latency, real-time responsiveness, and privacy protection. The decentralized approach supported by these solutions leverages our hardware platforms, AI accelerators and software frameworks to perform AI inference tasks locally on edge devices, reducing reliance on cloud-based services and enabling autonomous decision-making in distributed systems. Visit renesas.com/AI to put the Renesas approach to work on your next intelligent design.

Reference Materials

Articles and white papers:

Design AI / ML Applications the Easy Way

Endpoint Intelligence

The Role of AI and Endpoint Real-time Data Analytics

AI as a Service for Signal Processing

How to Maximize the Lifespan of Electric Motors

News:

Renesas Extends AIoT Leadership with Integration of Reality AI Tools and e² studio IDE

Tools:

Reality AI Tools

e² studio

DRP-AI TVM

AI Navigator

RZ Application Zoo

RZ AI SDK

AI Partner Solutions

AIZIP: Person Access System, Defect Detection, Glass Break Detection, Speaker ID

Cyberon: DSpotter

Irida Labs: QR Scanning

NotaAi: Nota DMS

Plumerai: People Detection

Edge Impulse: Object Detection

Ignitarium: Human Pose Detection and Classification

Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu, Koto-ku, Tokyo 135-0061, Japan

https://www.renesas.com

Contact Information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit:

https://www.renesas.com/contact-us

Trademarks

Doc Number: R11WP0005EU0100