Rev. Roum. Sci. Techn.– Électrotechn. et Énerg.
Vol. **69**, *1*, pp. 97–102, Bucarest, 2024

INTERNET OF THINGS PLATFORM BENCHMARK: AN ARTIFICIAL INTELLIGENCE ASSESSMENT

ROBERT-ALEXANDRU CRACIUN, RADU-NICOLAE PIETRARU, MIHNEA ALEXANDRU MOISESCU

Keywords: Internet of things (IoT); Artificial Intelligence (AI); Hardware; Platforms; Single board computer (SBC); Benchmark.

The Internet of Things (IoT) represents a transformative technological concept that seamlessly becomes a part of the Internet across all industries. Artificial intelligence (AI) provides IoT with new capabilities used to analyze data in real-time and make informed decisions. There is a wide array of IoT devices with different computational capabilities and AI accelerators that need to be compared. The current paper proposes a comparison between two single-board computers using an existing AI benchmark.

1. **INTRODUCTION**

The Internet of Things (IoT) represents a transformative technological concept in which everyday objects that all people use will be equipped with a blend of hardware and software components that will enable devices to communicate with each other and seamlessly become a part of the Internet. Through facilitating accessibility and engagement with a diverse array of devices such as home appliances, surveillance cameras, vehicles, *etc*., IoT is set to catalyze the development of a wide range of applications [1]. With the continuous integration of IoT technologies, they are leading the way into a new era of innovation, automation, and efficiency across a wide range of domains, having the potential to improve our lives and revolutionize industries.

IoT systems have the flexibility to be deployed in a wide range of industries. Each solution can be customized to suit specific needs and requirements for each applicability domain. In the realm of smart homes [2], IoT technologies provide users with the ability to remote control and monitor their appliances, security systems, and environmental settings. In the industrial sector [3] IoT is used to predict maintenance needs and improve machines and processes, leading to reduced downtime and lower operational expenses. Smart cities [4] leverage IoT to improve urban infrastructure, traffic management, waste disposal, and energy consumption. In the field of agriculture [5] IoT allows farmers to keep an eye on the crop conditions and adopt data-driven decisions to enhance the quality of the crops. In the automotive sector [6], the focus on connected vehicles enabled them to communicate with other vehicles and with the infrastructure, resulting in enhanced safety, traffic management, and improved user experience. With their wide range and flexible applications, IoT devices represent the main technology that is reshaping our world by enhancing the efficiency of everyday tasks in each industry, fostering innovation, and enhancing the overall quality of life. All those systems evolved into a significant data source with billions of interconnected devices that produce huge amounts of data that need to be analyzed [7], which can be addressed by leveraging artificial intelligence techniques.

The IoT data streams are continuously generating substantial heterogeneous quantities of data that are not practical to analyze in real-time [8]. Artificial intelligence (AI) techniques provide IoT devices with the capability to analyze data in real-time, make informed decisions, and establish automation processes driven by these algorithms. When developing AI-driven solutions for IoT, it is essential to consider the platform on which the application is deployed.

IoT includes a wide array of devices with different computational capabilities. If we are considering the AI domain, there are two broad categories: traditional devices with CPUs and IoT devices equipped with AI accelerators. Traditional IoT devices are typically powered by a CPU which can provide the user with a variety of tasks that can be handled by the device. These devices are suited for basic data processing, sensor monitoring, and communicating with other systems. In contrast, IoT devices equipped with AI accelerators, such as GPUs (graphics processing units) or AI chips like TPU (tensor processing units) or NPU (neural processing units) are designed to perform AI tasks making it possible for the IoT devices to analyze data in real-time and provide advanced features like object recognition, natural language processing or predictive analytics.

The current paper will propose a benchmark comparison between a traditional IoT device and an AI-powered IoT device to evaluate the performances of each system and determine which platform can be used to build a robust IoT security gateway that will also have advanced AI capabilities for threat detection and mitigation. This paper will contain a list of currently available benchmark suites, an overview of the chosen benchmark and the tests that will be performed on each platform, a discussion on the results, and how the chosen platform will be used to further develop the secure IoT gateway.

2. **RELATED WORK**

AI benchmarks are tools that are used in the landscape of AI research and development. They are used as objective and standardized methods for evaluating the performance of AI systems, models, and hardware across a diverse array of tasks. This chapter will focus on various research papers that offer unique insights into AI benchmarking, offer clarity on the scope and implementation of the benchmarking, and provide perspectives and contributions. Each paper presents a distinct viewpoint on the challenges and opportunities within AI benchmarking, and they will be used to deepen the understanding of how they can be used to properly assess different IoT hardware platforms.

AI benchmark [9] is a versatile benchmarking suite that proves the evaluation of AI systems and hardware performance. It covers a variety of diverse AI workloads, ranging from image recognition and object detection to speech recognition, ensuring a thorough assessment of AI capabilities. Utilizing well-known neural network models like MobileNet, Inception, ResNet, and BERT, it measures the efficiency with which the systems handle the model architectures in AI applications. AI Benchmark is adaptable across various platforms, making it suitable for assessing AI performance for devices like smartphones, edge

National University of Science and Technology "Politehnica" of Bucharest, Romania. Emails: robert.craciun2811@upb.ro, radu.pietraru@upb.ro, mihnea.moisescu@upb.ro

DOI: 10.59277/RRST-EE.2024.69.1.17

devices, and datacenter servers. As an open-source tool, AI Benchmark helps users make informed decisions regarding their hardware and software configurations for AI applications. Compared to other benchmarks, AI Benchmark excels in the variety of models tested compared to the upcoming benchmarks that are more algorithm specific.

MLPerf [10] represents an industry-standard benchmark suite dedicated to assessing the performance of machine learning models and hardware. The benchmark offers a comprehensive range of tasks encompassing image and speech recognition, as well as natural language processing. The importance of MLPerf is amplified by its extensive embrace within the AI community. What sets MLPerf apart is its commitment to delivering clear and uniform metrics, facilitating substantial cross-platform contrasts. Compared to AI Benchmark, the metrics produced after each test are more comprehensive and can be used to compare additional characteristics for different platforms.

AI matrix [11] is an evaluation and testing framework developed by Alibaba Group to assess the performance and capabilities of AI systems. This benchmark is designed to provide assessments of AI models across various domains, including natural language processing, computer vision, and more. By offering a standardized set of tests and metrics, this benchmark allows users to fine-tune their AI models by the three types of benchmarks proposed: layer-based, which consists of commonly used layers in deep learning neural networks like convolutional layer, activation layer, *etc*., macro benchmark, which consists of full models used in deep learning neural networks, and micro-benchmark, which consists of matrix operations. Compared to the other benchmarks, the current benchmark has an emphasis on deep learning algorithms, and it may not offer a complete comparison between platforms.

Fathom [12] offers a collection of models that incorporate multiple layer types, such as convolutional, fully connected, and recurrent neural network layers. However, its primary emphasis is on maximizing throughput rather than achieving the highest level of accuracy. This benchmark behaves like an AI benchmark that has different tests that don't focus on accuracy but on a variety of models tested.

EEMBC MLMark [13] is an AI benchmark designed to assess the machine learning inference performance of edge devices. It provides a comprehensive evaluation of the effectiveness of which a system can execute machine learning workloads, covering a wide range of neural network models and operations that include image classification and object-detection tasks. The benchmark measures aspects such as inference speed, power consumption, and memory usage. Like the AI matrix, this benchmark focuses on specific types of AI models and not on a wider array of models.

DAWNBench [14] was the first benchmark suite designed to evaluate and compare end-to-end performance of deep learning training across various machine learning models and hardware configurations. It offers a diverse set of tasks, including image classification and language modeling that allows users to assess the speed and efficiency of AI models on different platforms. This benchmark has an emphasis on deep learning algorithms and compared to AI Benchmark or Fathom which have a wider range of tests available, it may not offer the best results for platform comparison.

DeepBench [15] is a microbenchmark specialized in assessing the low-level performance of both hardware and software components in the context of deep learning workloads. It focuses on quantifying the computational efficiency of fundamental operations crucial for deep learning, such as convolutions and matrix multiplications that are made at the kernel level. This benchmark focuses on low-level operation analysis and its granularity may not be suitable for testing overall platform AI capabilities.

AI benchmarks play an important role in the dynamic world of artificial intelligence. They provide a standardized and objective way to assess the performance of AI systems, models, and hardware across a multitude of tasks, ensuring that technological advancements are rigorously evaluated and compared. Based on the current research, in the next chapters will assess how these benchmarks can facilitate the selection of a hardware IoT platform for the development of a highly secure IoT gateway.

3. TRADITIONAL VS. AI-READY IOT PLATFORMS

With the integration of AI, IoT platforms have experienced a significant evolution in their applicability scope. This transformation enabled the IoT platforms to perform complex AI tasks, which are capable of handling and processing vast amounts of data while making intelligent decisions in real-time. To properly compare the currently available IoT platforms, there needs to be a comparison between hardware components, processing capabilities, and AI-specific technologies like neural processing units (NPU), tensor processing units (TPU), and graphics processing units (GPUs).

Traditional IoT platforms are conventionally using microcontrollers, sensors, and communication modules. These devices are designed to collect and transmit data without any significant processing power for AI activities. Traditionally, data processing is being done on external cloud resources or edge computing devices, which can lead to latency and increased data transfer costs. Even though basic AI analytics can be performed on these types of devices, the platforms are not equipped with capabilities for complex AI models for real-time analysis. On the other hand, AI-ready platforms use more powerful acceleration processors such as GPUs, TPUs, and NPUs that allow the platforms to efficiently handle AI workloads and accelerate AI model development and their inferencing processes. These platforms can handle data processing both on the edge and in the cloud offering flexibility in terms of latency and cost since they can filter and aggregate data locally and transmit only relevant insights to the cloud. The computation of the information is highlighted by the complex AI capabilities that allow the execution of AI models at the edge making real-time decisions possible. Acceleration processors provide different capabilities based on their processor type.

NPUs [16] are specialized processors designed for a wide range of AI application scenarios. Compared to traditional CPU and GPU computations, NPUs are more power efficient and deliver enhanced performance in AI inference, gaining an edge in the realm of low-powered IoT devices. On the other hand, GPUs [17] are suitable for training deep learning models, but they can use more power than NPUs and TPUs, which makes them less efficient for AI IoT scenarios. Lastly, TPUs [18] are a technology developed by Google [19] offering similar capabilities to the NPUs, but it is more focused on deep learning tasks and may have limited flexibility for other AI applications.

Out of all the presented options, the best IoT candidate for AI is the NPU accelerator since it offers the most balanced characteristics for a wide range of IoT applications. It is built with a focus on power efficiency, making it suitable for battery-

powered IoT devices as they consume less energy and deliver fast and efficient AI inference, ensuring real-time decision-making and reduced latency for IoT solutions. The power efficiency nature of the NPUs can lead to cost savings and offer robust performances for the IoT environments.

AI-ready IoT platforms equipped with NPU accelerators are transforming IoT solutions by enabling real-time AI processing and enabling new capabilities for everyday use cases. NPUs are a compelling option for IoT applications due to efficient AI performance with low energy usage, giving them an edge over traditional CPU and GPU platforms. However, there is a need to conduct a performance comparison between traditional IoT devices and NPU accelerated devices to validate if the latest is more efficient for AI IoT applications.

This validation is necessary to determine the platform's suitability for building a performant IoT security gateway that will leverage AI algorithms for intrusion detection and intelligent resolution to solve security issues.

4. **METHODOLOGY**

According to the literature reviewed in this paper, considering that various benchmarks can be used to compare different hardware platforms, it is essential to select a specific benchmark that will be used for a comparison that aims to identify the most suitable platform for AI applications, particularly to be used for an IoT security gateway with AI capabilities.

While all the benchmarks examined in this paper are powerful options for evaluating and comparing AI capabilities across various hardware platforms, the choice was made for the AI benchmark. This benchmark was chosen because it covers a more diverse array of tasks that provides a broad spectrum of AI workloads making it suitable to assess the overall performance of the selected platforms. As the scope of the current paper is set to find a suitable platform that can be used to develop an IoT cybersecurity gateway with AI capabilities, the scope is to find the best platform that performs better in different scenarios, with different models and different datasets. AI Benchmark has an edge over the other benchmarks as it offers several test cases, with different models and datasets that can be assessed. The decision was also made to focus on SBC platforms which will be presented later in this chapter. However, the other benchmarks still offer robust solutions to compare platforms, but they weren't chosen for various reasons. MLPerf is known for its complexity which may be excessive for the relative constraint environments as the used datasets are not suitable for SBCs as they have several GB worth of data, AI matrix, and DAWNBench have an emphasis on deep learning tasks and are not testing different algorithms, Fantom is designed for Intel platforms, while the SBC use ARM CPUs, EEMBC MLMark is tailored for microcontroller platforms and DeepBench is designed for low-level operation analysis.

AI benchmark consists of 19 sections that contain a total of 42 tests as presented in Table 1.

The benchmark will produce different timing metrics that will be used to compare the platforms: benchmark execution time, inference time, training time, AI score, inference score, and training score. Those scores represent a mean comparison between the reference time that each algorithm should take to execute the process and the mean times that each algorithm had. The total AI score is the summation of the inference and training scores. The platforms that will be assessed are presented in Table 2.

The benchmark results that will be assessed in the following chapter will give enough information to make a decision on which platform is best suited to host a powerful security IoT gateway with AI capabilities.

Table 1

AI benchmark test

Model	Type	Paper
MobileNet-V2	Classification	[20]
Inception-V3	Classification	[21]
Inception-V4	Classification	[22]
Inception-ResNet-V2	Classification	[23]
ResNet-V2-50	Classification	[23]
ResNet-V2-152	Classification	[23]
VGG-16	Classification	[24]
SRCNN 9-5-5	Image-to-Image Mapping	[25]
VGG-19	Image-to-Image Mapping	[26]
ResNet-SRGAN	Image-to-Image Mapping	[27]
ResNet-DPED	Image-to-Image Mapping	[28]
U-Net	Image-to-Image Mapping	[29]
Nvidia-SPADE	Image-to-Image Mapping	[30]
ICNet	Image Segmentation	[31]
PSPNet	Image Segmentation	[32]
DeepLab	Image Segmentation	[33]
Pixel-RNN	Image Inpainting	[34]
LSTM	Sentence Sentiment Analysis	[35]
GNMT	Text Translation	[36]

Table 2

Tested Platforms Specifications

Characteristic	Raspberry Pi 5	Orange Pi 5
Socket	BCM2712	RK3588S
CPU	ARM-Cortex A76	ARM-Cortex A76 + ARM-Cortex A55
RAM	4GB LPDDR4X	4GB LPDDR4/4x
Storage	SD Card	SD Card
External Storage	USB SSD SATA3	USB SSD SATA3
AI Accelerator	N/A	NPU Accelerator – 6TOPS

5. **RESULTS**

The AI benchmark was executed on both platforms to compare the performance of each hardware configuration. The scope of the current study is to find which platform can be used as a security gateway with performant AI capabilities for IoT devices. The performance is measured by the following metrics: benchmark execution time, inference time, training time, AI score, inference score, and training score.

1. Benchmark execution time. Table 3 contains a comparison between the total execution time for a benchmark run for each platform.

Table 3

Benchmark execution time [s]

Platform	Benchmark execution time [s]
Raspberry Pi 5	4377
Orange Pi 5	3570

The execution time results are as expected. The Orange Pi 5 has a more powerful hardware configuration than the Raspberry Pi 5. Considering the total time execution of the benchmark, the Orange Pi 5 demonstrates almost 20 % greater time efficiency compared to the Raspberry Pi 5. The impact of this metric will be reflected in the inference and training times reflected in the upcoming benchmark results.

2. Inference mean time. Figure 2 and Table 4 compare all the benchmark inference tests. Out of 19 categories,

the Orange Pi 5 had an edge over the Raspberry Pi 5 in 16 categories. The results are expected since the Orange Pi 5 has a more performant CPU and an NPU that work together in the inference process. However, the other three categories need to be better optimized to use the full potential of the Orange Pi 5 board, and the Raspberry Pi 5 has better results on those tests. The CPU of the Orange Pi 5 will also provide advantages in the model training over the Raspberry Pi 5.

Table 4

Inference mean time [s]

Benchmark Test	Raspberry Pi 5	Orange Pi 5
MobileNet-V2	2.601	1.401
Inception-V3	5.734	3.85
Inception-V4	5.762	3.972
Inception-ResNet-V2	6.728	5.114
ResNet-V2-50	3.745	2.982
ResNet-V2-152	5.589	4.183
VGG-16	9.723	8.067
SRCNN 9-5-5	7.866	6.066
VGG-19 Super-Res	13.517	10.124
ResNet-SRGAN	14.675	11.288
ResNet-DPED	15.697	11.164
U-Net	34.212	26.07
Nvidia-SPADE	13.089	10.128
ICNet	5.464	9.416
PSPNet	65.432	49.862
DeepLab	13.967	14.57
Pixel-RNN	8.656	5.916
LSTM-Sentiment	22.475	26.015
GNMT-Translation	7.2	6.949

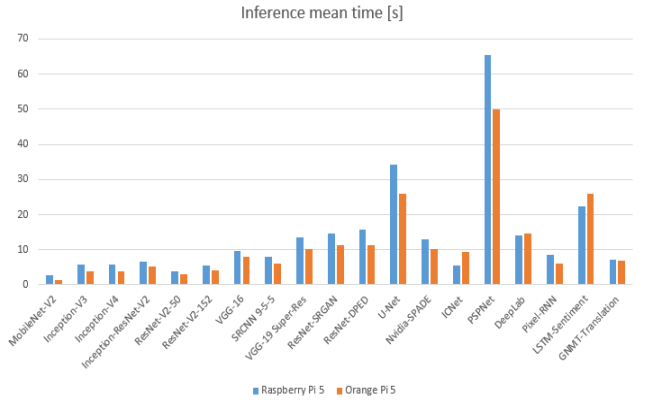

Fig. 2 – Inference mean time [s].

3. Training mean time. Figure 3 and Table 5 compare benchmark training tests. The benchmark skipped the GMMT algorithm as it used a pre-trained model. Of the 19 tests, the Raspberry Pi 5 had an edge only on the LSTM algorithm, which trained the model faster than the Orange Pi 5. In all the other scenarios, Orange Pi 5 had better training times versus the Raspberry Pi 5 and the most notable difference is on the SRCNN 9-5-5 model where Orange Pi 5 had 72 % more-time performance than the Raspberry Pi 5.

4. AI score. Figure 4. contains a comparison between the AI scores that the AI Benchmark produced. The total score is the summarization of the mean inference time score and mean training time score. The scores represent a comparison between each algorithm reference time and the actual mean time for inference, respectively for the training times. The times presented in the last two categories produced the scores and based on them, the winner in this category is the Orange Pi 5 which has an AI score of 285 points, which is 32 % more performant than the result of the Raspberry Pi 5.

Table 5

Training mean time [s]

Benchmark Test	Raspberry Pi 5	Orange Pi 5
MobileNet-V2	17.633	8.594
Inception-V3	40.623	24.668
Inception-V4	47.731	24.21
Inception-ResNet-V2	44.507	21.783
ResNet-V2-50	38.824	13.746
ResNet-V2-152	43.876	21.132
VGG-16	34.617	15.798
SRCNN 9-5-5	465.584	129.051
VGG-19 Super-Res	61.895	57.048
ResNet-SRGAN	40.668	33.05
ResNet-DPED	90.241	39.165
U-Net	74.303	35.112
Nvidia-SPADE	44.23	15.061
ICNet	33.386	18.074
PSPNet	43.588	29.592
DeepLab	23.692	14.229
Pixel-RNN	5.729	4.262
LSTM-Sentiment	31.713	57.607
GNMT-Translation	0	0

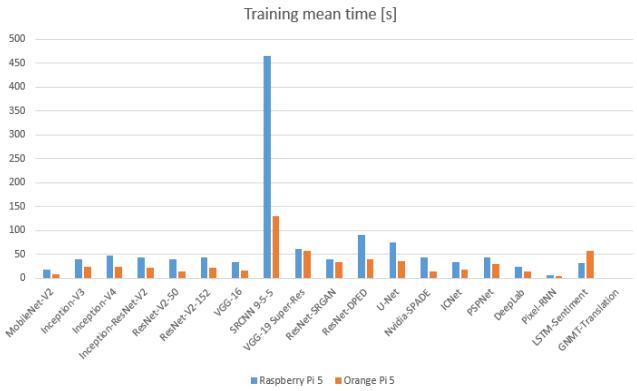

Fig. 3 – Training mean time [s].

5. Platform benchmark conclusions. Both platforms offer robust capabilities for standard tasks but are also good options for AI tasks. The AI Benchmark offers a good suite of tests that were used to compare the Raspberry Pi 5 and Orange Pi 5 platforms to determine which one is the better to be used for a secure IoT gateway with AI capabilities. Even though the Raspberry Pi 5 has robust results in the AI Benchmark and can be used for a secure IoT gateway with AI capabilities, also considering the price point and energy efficiency of the SBC, the Orange Pi 5 represents a better option for a small cost increase and similar energy performance has better results in almost all aspects tested using the AI Benchmark tool.

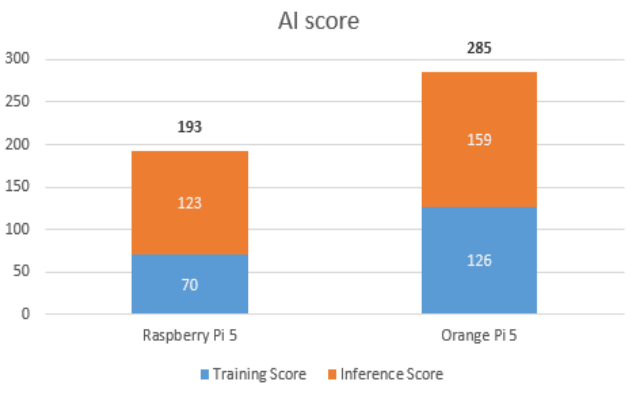

Fig. 4 – AI score.

6. FUTURE DEVELOPMENT

The conducted AI Benchmark comparison was made to identify which of the two platforms can be used to create a powerful identification and prevention machine learning system for IoT devices that communicate using the TCP/IP protocol. The Orange Pi 5 will be used to create a system that will capture traffic data, create a new dataset to train ML models for device identification and threat detection, and do real-time analysis and response to protect the connected IoT devices in the system. The benchmark execution was limited to using CPU only as it wasn't designed to leverage the onboard NPU of the Orange Pi 5. Even though only the CPU was used for the comparison, the benchmark offers important results as the machine learning models need to be constantly retrained to detect newly added devices in the system and new types of vulnerabilities. In the future, enabling the NPU on the Orange Pi 5 will ensure more efficient inferencing for the created machine learning models.

In the continuously expanding field of IoT, where security is one of the most important factors in IoT ecosystems, exploring the potential of the NPU becomes mandatory. While the CPUs show good performance, AI accelerators offer greater computational efficiency and real-time detection and response for external threats, offering robust protection mechanisms. The future strategy is to minimize the vulnerabilities and create a safe environment for IoT devices which must be used securely, contributing to the development of IoT.

7. CONCLUSIONS

The current study has undertaken a thorough comparison of two platforms, the Raspberry Pi 5 and Orange Pi 5, to determine which platform is best suited to be used as an intelligent IoT security gateway with AI capabilities. The evaluation was facilitated by the comprehensive suite of tests provided by AI Benchmark, which allowed a thorough investigation of the AI capabilities of each platform. The results showed that the Orange Pi 5 can be successfully used to create a robust security IoT gateway with AI capabilities for IoT devices that communicate using the TCP/IP protocol. As the demand for efficient and secure IoT environments continues to grow, the insights collected from this comparative analysis provide a valuable foundation for the decision made in selecting the platform for future IoT implementation.

Received on 26 November 2023

REFERENCES

1. A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, *Internet of Things for Smart Cities*, IEEE Internet of Things Journal, **1**, *1*, pp. 22–32 (Feb. 2014).
2. A.C. Gheorghe, H. Andrei, E. Diaconu, G.C Seritan, B. Enache, *smart system for standby power consumption reduction of household equipment,* Rev. Roum. Sci. Techn. – Électrotechn. et Énerg., **68**, *4*, p. 413–418 (2023).
3. M. Soori, B. Arezoo, R. Dastres, *Internet of things for smart factories in industry 4.0, a review*, Internet of Things and Cyber-Physical Systems, **3**, pp. 192–204 (2023).
4. A. Rejeb, K. Rejeb, S. Simske, H. Treiblmaier, S. Zailani, *The big picture on the internet of things and the smart city: a review of what we know and what we need to know*, Internet of Things, **19**, 100565 (2022).
5. M. Dhanaraju, P. Chenniappan, K. Ramalingam, S. Pazhanivelan, R. Kaliaperumal. *Smart Farming: Internet of Things (IoT)-Based Sustainable Agriculture*, Agriculture, **12**, *10*, pp. 1745 (2022).
6. S.J. Muthiya, S. Anaimuthu, J.A. Dhanraj, N. Selvaraju, G. Manikanta, C. Dineshkumar *Application of Internet of Things (IoT) in the Automotive Industry*. Integration of Mechanical and Manufacturing Engineering with IoT (eds R. Rajasekar, C. Moganapriya, P. Sathish Kumar, M. Harikrishna Kumar), 2023.
7. A. Nauman, Y.A. Qadri, M. Amjad, Y.B. Zikria, M.K. Afzal, S.W. Kim, *Multimedia Internet of Things: a comprehensive survey*, IEEE Access, **8**, pp. 8202–8250 (2020).
8. G. Yashodha, P.R. Pameela Rani, A. Lavanya, V. Sathyavathy *Role of Artificial Intelligence in the Internet of Things – A Review*, IOP Conference Series: Materials Science and Engineering, **1055** (2021).
9. A. Ignatov, R. Timofte, A. Kulik, S. Yang, K. Wang, F. Baum, M. Wu, L. Xu, L. Van Gool, *Ai benchmark: All about deep learning on smartphones in 2019*, arXiv preprint arXiv:1910.06663 (2019).
10. V. Reddi *et al*., *Mlperf inference benchmark*, 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), pp. 446–459 (2019).
11. Alibaba, *Ai matrix* https://aimatrix.ai/en-us/, Alibaba, 2018.
12. R. Adolf, S. Rama, B. Reagen, G.Y. Wei, D. Brooks, *Fathom: reference workloads for modern deep learning methods* IEEE, International Symposium on Workload Characterization (IISWC), 2016.
13. EEMBC, *Introducing the EEMBC MLMARK benchmark* https:// www.eembc.org/mlmark/index.php, Embedded Microprocessor Benchmark Consortium, 2023.
14. C. Coleman, D. Narayanan, D. Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. R´e, M. Zaharia, *DAWNBench: an end-to-end deep learning benchmark and competition*, NeurIPS ML Systems Workshop, 2017.
15. Baidu-research, *DeepBench: benchmarking deep learning operations on different hardware*, https://github.com/baidu-research/DeepBench (2023).
16. T. Tan, G. Cao, *FastVA: deep learning video analytics through edge processing and NPU*, Mobile IEEE INFOCOM, IEEE Conference on Computer Communications, Toronto, ON, Canada, pp. 1947–1956, 2020.
17. X. Qi, C. Liu, *Enabling deep learning on IoT edge: approaches and evaluation*, IEEE/ACM Symposium on Edge Computing (SEC), Seattle, WA, USA, pp. 367–372, 2018.
18. A. Shahid, M. Mushtaq, *A survey comparing specialized hardware and evolution*, TPUs for Neural Networks, 23rd IEEE International Multitopic Conference (INMIC), Bahawalpur, Pakistan, pp. 1–6, 2020.
19. Google, *Cloud tensor processing units (TPUs)*, https://cloud.google. com/tpu, 2023.
20. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, *MobileNets: efficient convolutional neural networks for mobile vision applications*. ArXiv, abs/1704.04861 (2017).
21. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*, IEEE Conference on Computer Vision, and Pattern Recognition (CVPR), pp. 2818–2826, 2016.
22. C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, *Inception-v4, Inception-ResNet and the impact of residual connections on learning*, 2016, ArXiv, abs/1602.07261.
23. K. He, X. Zhang, S. Ren, J. Sun, *Identity mappings in deep residual networks*, European Conference on Computer Vision, 2016.
24. K. Simonyan, A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, CoRR, abs/1409.1556, 2014.
25. C. Dong, C.C. Loy, K. He, X. Tang. *Image Super-Resolution Using Deep Convolutional Networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **38**, pp. 295–307 (2014).
26. J. Kim, J.K. Lee, K.M. Lee., *Accurate image super-resolution using very deep convolutional networks*, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654, 2016.
27. C. Ledig, L. Theis, F. Huszár, J. Caballero, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, *Photo-realistic single image super-resolution using a generative adversarial network*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114, 2017.
28. A.D. Ignatov, N. Kobyshev, K. Vanhoey, R. Timofte, L.V. Gool, *DSLR-quality photos on mobile devices with deep convolutional networks*. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3297–3320, 2017.
29. O. Ronneberger, P. Fischer, T. Brox, *U-Net: convolutional networks for biomedical image segmentation*, Navab, N., Hornegger, J., Wells, W., Frangi, A. Eds., Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science, **9351**, Springer, Cham (2015).

30. T. Park, M.Y. Liu, T.C. Wang, J.Y. Zhu, *semantic image synthesis with spatially-adaptive normalization*, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 2332–2341 (2019).
31. H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, *ICNet for real-time semantic segmentation on high-resolution images*, V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss Eds., Computer Vision – ECCV 2018. Lecture Notes in Computer Science, **11207**, Springer, Cham, 2018.
32. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, *Pyramid scene parsing network*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239, 2016.
33. G. Papandreou, L. Chen, K.P. Murphy, A.L. Yuille, *Weakly- and semi-supervised learning of a DCNN for semantic image segmentation*. ArXiv, abs/1502.02734, 2015.
34. A.V. Oord, N. Kalchbrenner, K. Kavukcuoglu, *Pixel recurrent neural networks*, International Conference on Machine Learning, 2016.
35. S. Hochreiter, J. Schmidhuber, *Long short-term memory*, Neural computation, **9**, pp. 1735-80, 1997, 10.1162/neco.1997.9.8.1735.
36. Y. Wu et al., *Google's neural machine translation system: bridging the gap between human and machine translation*, ArXiv, abs/1609. 08144, 2016.