

AI PCs and On-Device Artificial Intelligence: The Future of Intelligent Computing

Sasane Sakshi Shirish, Prof.Bhosale S.B, Prof.Bangar A.P, Prof.Khatrri V.A

Department of Computer Engineering
Jaihind college of engineering SPPU, Pune,India

Abstract- AI PCs and on-device artificial intelligence are changing how everyday computers use AI. Instead of sending every request to a remote cloud server, a device can process suitable tasks locally using its CPU, GPU, and specialised neural processing unit (NPU). This can reduce response time, limit data transfer, support some functions without a stable internet connection, and give users more control over personal data. The practical value of these benefits depends on the task, the model, and the computer's hardware. Local processing also has to work within limits such as memory capacity, power consumption, heat, and model size. This review discusses three research works: a comprehensive survey of on-device AI models, a study proposing a collaborative consumer edge-AI ecosystem, and research on on-device AI for smart environments and Industry 4.0. The papers are compared by focus, approach, contribution, and limitation. The review identifies model optimisation, efficient hardware use, privacy-aware collaboration, and energy management as important areas. It also finds that the selected works do not establish a common benchmark specifically for AI PCs. A task-aware local/cloud architecture is proposed for future study; it has not been experimentally tested.

Keywords— AI PCs; On-Device AI; Edge Computing; NPU; Model Optimisation; Privacy; Energy Efficiency; Hardware Acceleration.

I. INTRODUCTION

Artificial intelligence is becoming part of common computing activities such as writing, searching, summarising documents, editing images, translation, and programming assistance. These features are gradually moving from specialised applications into the everyday tools people use for study, work, communication, and creative tasks. As a result, the way a computer processes AI requests has become an important part of the user experience.

Many AI services traditionally depend on cloud computing. In this arrangement, a request is sent through the internet to a remote server, processed there, and returned to the user's device. Cloud systems can provide access to powerful models and large amounts of computing capacity. However, the experience may depend on network quality, server availability, service policies, and the time needed to transfer information. Sending personal or work-related information to an external service can also raise privacy and security questions.

On-device AI takes a different approach: an AI model runs directly on the device that receives the data. In an AI PC, the central processing unit (CPU), graphics processing unit (GPU), and neural processing unit (NPU) may work together. The CPU handles general computing and coordinates tasks, the GPU is suited to parallel workloads, and an NPU accelerates supported AI operations efficiently. The exact capabilities depend on the computer's hardware, drivers, operating system, and available software.

Local AI can provide quicker responses, reduce dependence on internet access, and keep some information on the computer. These benefits are not automatic for every task. Large models may still need cloud resources, while local processing is limited by memory, power, heat, and model accuracy. A practical system must therefore decide which tasks are suitable for local execution and when remote processing may be useful. Wang et al. [1] survey model optimisation, Laskaridis et al. [2] discuss collaboration among consumer devices, and De Vita et al. [3] examine on-device AI in smart and industrial environments. This review compares these works and identifies directions relevant to AI PCs.

II. REVIEW METHODOLOGY AND LITERATURE SELECTION CRITERIA

This is a qualitative, literature-based review of three supplied research papers. The papers were selected because they cover complementary aspects of on-device AI: model design and optimisation [1], a collaborative consumer edge-AI ecosystem [2], and applications and system requirements in smart environments and Industry 4.0 [3]. Together, they offer a useful overview of the topic from the level of an individual model to the coordination of devices and the requirements of real-world applications.

Each paper was examined for its purpose, technical approach, main contribution, limitations, and future research needs. The information was grouped under model efficiency, hardware coordination, privacy, energy use, and applications. Particular attention was paid to whether the discussion could be applied to AI PCs, while recognising that some findings concern phones, embedded devices, or industrial systems rather than personal computers specifically. Because the review is limited to three papers and the studies have different scopes, it is a focused narrative review rather than an exhaustive systematic review. Their results are compared qualitatively, not as directly equivalent performance measurements. No new experiments or benchmark tests were conducted for this review, so the discussion identifies potential implications rather than claiming measured performance improvements.

III. LITERATURE REVIEW

1. Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models

Wang et al. [1] survey the concepts, applications, challenges, and optimisation methods associated with AI models running on edge devices. The central idea is to bring at least part of the intelligence closer to the source of the data instead of relying on a remote server for every request. Local AI can reduce reliance on cloud infrastructure and support timely processing, especially when data is sensitive or a quick response is required. The paper covers applications including mobile devices, smart

systems, autonomous technologies, and medical devices.

A key contribution is its review of techniques for fitting AI models to limited hardware. Model compression reduces the overall size of a model; pruning removes parameters or connections that contribute less to the result; and quantisation uses lower numerical precision to reduce memory demand and improve inference speed on supported hardware. These approaches can make deployment more practical, but they need to be selected carefully because a smaller model may not preserve exactly the same output quality in every task.

Hardware acceleration and hardware–software co-design can further improve efficiency. In practical terms, the model and the processor should be considered together rather than designed as separate parts. This is relevant to AI PCs because local assistants and generative AI models must operate within available memory and thermal limits. However, the survey covers many types of edge devices rather than AI PCs alone, and the benefit of each technique depends on the model, hardware, and task. It therefore provides useful principles but does not by itself establish how every AI PC will perform.

2. The Future of Consumer Edge-AI Computing

Laskaridis et al. [2] argue that consumer devices could move beyond working as separate systems. Their proposed EdgeAI-Hub concept coordinates AI tasks and available resources across nearby devices, with optional cloud support. The paper highlights shared computation, shared context, privacy, and sustainability. This perspective is useful because people increasingly use several connected devices in the same environment, and those devices often have different capabilities.

For example, a home may contain a laptop, phone, television, and other AI-capable devices with different computing power and availability. Coordinating resources may help distribute workloads instead of placing every task on one device. A capable device could potentially help with a task when another device is busy or has fewer

resources. Such coordination might also allow a system to use resources more flexibly, rather than assuming that all processing must take place on a single computer or in a remote data centre.

The approach also creates challenges. Devices may become unavailable, have different hardware, or belong to different users. Sharing data requires clear permissions and trust boundaries, and the system needs a reliable way to respond if a device disconnects during a task. Compatibility between operating systems and software platforms can also affect whether collaboration works smoothly. This work is a proposed ecosystem and design direction, not a feature guaranteed to be present in all consumer computers. Still, it provides a useful way to think about future AI PCs that may collaborate with other trusted devices or use cloud resources only when needed.

3. On-device Artificial Intelligence Solutions with Applications to Smart Environments

De Vita et al. [3] present research on on-device AI for smart environments and Industry 4.0. In these settings, devices sense conditions and make decisions close to where data is produced. Examples of such environments include connected systems that monitor activity, equipment, or surrounding conditions and then respond based on the information collected. Local processing can reduce communication overhead and response time, while limiting the need to send sensitive information to remote servers.

The paper also emphasises that embedded and edge devices have limited memory, processing power, and energy, which restricts the size and complexity of models that can run efficiently.

A solution that performs well in a laboratory may need further optimisation before it can operate continuously in a real environment. Reliability matters as well: the system should continue to behave predictably when connectivity is weak or the available resources change.

The work highlights the need for appropriate architectures, frameworks, and privacy and security

protections. These principles also apply to AI PCs, although a PC usually has more resources than a small embedded device. A computer still needs efficient models and sensible power use, especially when AI tasks run alongside everyday applications. Since the paper focuses on smart environments and industrial applications, it provides relevant background rather than direct performance measurements for consumer AI PCs.

4. Synthesis of the Reviewed Studies

The three papers look at different levels of the same problem. Wang et al. [1] focus on efficient models, Laskaridis et al. [2] on collaboration between consumer devices, and De Vita et al. [3] on application needs and system constraints.

Together, they show that on-device AI depends not only on a capable processor but also on model size, software support, privacy, and energy management. The system must balance these factors instead of treating processor speed as the only measure of success.

Local inference should also be distinguished from local training. Inference means using a trained model to generate a prediction or response, while training involves learning or updating model parameters.

Running a trained model is generally less demanding than training or adapting it, and not every PC can perform every type of training efficiently. This distinction is important when describing what an AI PC can do in everyday use.

IV. COMPARATIVE ANALYSIS OF EXISTING STUDIES

Table I compares the focus and contribution of the selected works. Since their objectives differ, a single numerical ranking would not be meaningful. Instead, the comparison highlights how each paper contributes a different part of the overall picture and where further AIPC-specific testing is needed.

Table 1. Qualitative comparison of the three selected research works.

Study	Main focus	Contribution	Limitation for AI-PC use
Wang et al. [1]	On-device models and optimisation	Reviews compression, pruning, quantisation, hardware acceleration, and deployment challenges.	Broad edge-AI survey; not limited to AI PCs or one common benchmark.
Laskaridis et al. [2]	Consumer edge-AI ecosystem	Proposes shared compute/context, resource orchestration, privacy-aware collaboration, and sustainable resource use.	Conceptual ecosystem; coordination, trust, and compatibility need validation.
De Vita et al. [3]	Smart environments and Industry 4.0	Highlights local processing, application needs, resource limits, and privacy.	Focus is broader than consumer PCs; results vary by device and task.

1. Local AI Compared with Cloud AI

Local and cloud AI are not complete substitutes. Local processing can provide quick responses and keep selected data on the computer, but model size is limited by the machine's resources. Cloud processing can provide access to larger models, but it depends on connectivity and may involve transferring data to a service provider. A hybrid design can use local AI for suitable tasks and request cloud assistance only when necessary and permitted. The choice should depend on the type of request rather than on a fixed rule that all tasks must stay local or go to the cloud.

A short, routine request may be suitable for a local model, while a complex task may benefit from a larger remote model. Privacy settings, internet availability, response-time needs, and user permission should also be considered. A well-designed system should explain when information may leave the device and should not send sensitive content to an external service without an appropriate basis.

Table 2. General comparison of local and cloud-based AI processing.

Aspect	On-device processing	Cloud processing
Response time	Can be fast because data need not travel to a remote server.	Depends on network and server response.
Privacy	Data can remain local, but local security is still needed.	Data may be transferred under the provider's policies.
Resources	Limited by PC memory, processors, power, and cooling.	Can access larger remote computing resources.
Connectivity	Some functions can work offline.	Usually requires an internet connection.

V. PROPOSED CONCEPTUAL ARCHITECTURE

Based on the reviewed literature, a possible AI-PC architecture can use a task-aware processing pipeline. This is a conceptual proposal for future investigation; it has not been implemented or experimentally validated. The main aim is to make

processing decisions according to the needs of each task, available resources, and the user's privacy preferences.

Proposed flow: User request / AI task → Check task, privacy and resources → Local CPU / GPU / NPU or optional cloud processing → Result and resource monitoring.

The system receives a user request and checks the task type, privacy requirements, available memory, processor load, power conditions, and network availability. If the task fits the local resources, it is assigned to the CPU, GPU, NPU, or a combination supported by the system. The choice of processing unit should follow the application's capabilities and should avoid placing unnecessary load on one component.

If local execution is unsuitable, the system may offer cloud processing subject to user permission and the application's data policy. For example, a task may exceed the available memory or require capabilities that the local model does not provide. The system should not assume that cloud processing is always permitted; sensitive requests may need to remain local or be declined if the required level of protection cannot be met.

The final stage monitors output quality, response time, memory use, and energy use. These observations could help the system make better decisions for later tasks and identify when a model is too demanding for the computer. This design combines model optimisation [1], resource coordination and optional offloading [2], and privacy- and resource-aware application needs [3]. Since it is conceptual, its benefits must be tested through implementation and repeatable experiments before any performance claims can be made.

VI. RESEARCH CHALLENGES AND IDENTIFIED GAPS

1. Model Size, Memory and Performance

AI models can require substantial memory and computation. Compression, pruning, and

quantisation can reduce these demands, but their effect on accuracy depends on the model and task [1]. A model that is small enough to load may still respond slowly when the computer is running several applications at once. AI-PC studies therefore need repeatable tests across different hardware configurations and under realistic multitasking conditions. They should measure not only whether a model runs, but also whether its output remains useful and consistent.

2. Power Use and Thermal Limits

Local processing can reduce data transmission, but it does not automatically reduce total energy use. Continuous AI workloads can increase power consumption and heat, particularly in laptops where cooling capacity and battery life are important. If a processor reduces its speed to manage temperature, performance may change during a long task. Energy-aware scheduling should therefore be evaluated alongside response time and output quality [2], [3]. Tests should consider both short requests and longer periods of continuous use.

3. Privacy, Security and Compatibility

Keeping data local can reduce exposure during transmission, but applications, logs, permissions, and connected devices still need protection. A locally stored file or model can still be exposed if the computer or application is not secure. Collaboration introduces additional questions about consent and trust [2], particularly when data or intermediate results are shared between devices. Differences in CPUs, GPUs, NPUs, drivers, operating systems, and model formats can also produce inconsistent performance between PCs. Better compatibility testing and clear user controls would help address these concerns.

4. Lack of Common AI-PC Benchmarks

The selected works cover on-device AI broadly, consumer edge collaboration, and smart environments; they do not establish one common benchmark specifically for AI PCs. A useful benchmark should report task quality, latency, memory use, energy consumption, sustained performance, offline behaviour, and the amount of data sent outside the device. It should also describe

the model version, hardware configuration, software environment, and test conditions so that results from different systems can be compared fairly.

Table 3. Research gaps and possible future directions.

Research gap	Possible future direction
Results are difficult to compare across hardware and tasks.	Develop repeatable AI-PC benchmarks for quality, latency, memory and energy.
Local/cloud decisions are often application-specific.	Build task-aware routing based on privacy, connectivity, load and user choice.
Efficiency methods may affect output quality.	Measure accuracy and usability after compression, pruning and quantisation.
Cross-device collaboration raises trust issues.	Use consent-based collaboration, access controls and reliable fallback behaviour.

VII. CONCLUSION AND FUTURE SCOPE

AI PCs and on-device AI represent a shift towards running more AI tasks close to the user and the data. The reviewed literature shows that efficient models, suitable hardware, privacy-aware design, and resource management are all needed for practical deployment. Wang et al. [1] explain model optimisation; Laskaridis et al. [2] consider collaboration between consumer devices; and De Vita et al. [3] discuss on-device AI in smart and industrial environments. Although their areas of focus differ, each paper contributes to understanding the technical and practical conditions required for useful local AI.

The main benefit of local AI is not that cloud computing becomes unnecessary, but that some tasks can be completed with less delay and less data transfer. The main challenge is balancing output

quality with memory, power, temperature, compatibility, and privacy requirements. A task-aware hybrid design is therefore promising: use local processing when suitable and consider cloud support only when required and permitted. Such a design should keep users informed and give them meaningful control over how their information is processed.

Future work could implement this architecture on different AI-PC configurations and compare local, cloud, and hybrid execution using common tasks. Evaluation should include response time, quality, energy use, memory demand, sustained performance, offline behaviour, and data transfer. It would also be useful to test the effect of multitasking and compare model optimisation techniques under the same conditions. These steps could help move the discussion from general principles towards evidence that users and developers can apply when selecting or designing AI-PC systems.

REFERENCES

1. X. Wang, Z. Tang, J. Guo, T. Meng, C. Wang, T. Wang, and W. Jia, "Empowering Edge Intelligence: A Comprehensive Survey on OnDevice AIModels," ACM Computing Surveys, vol. 57, no. 9, Article 228, Apr. 2025, doi: 10.1145/3724420.
2. S. Laskaridis, S. I. Venieris, A. Kouris, R. Li, and N. D. Lane, "The Future of Consumer Edge-AI Computing," arXiv:2210.10514v3, Jun. 2024;extended version of a paper accepted at IEEE Pervasive Computing.
3. F. De Vita, D. Bruneo, and S. K. Das, "On-device Artificial Intelligence solutions with applications to smart environments," Future GenerationComputer Systems, vol. 180, article 108373, 2026, doi: 10.1016/j.future.2026.108373.
4. R. Wang, Z. Gao, L. Zhang, S. Yue, and Z. Gao, "Empowering Large Language Models to Edge Intelligence: A Survey of Edge-Efficient LLMs and Techniques," Computer Science Review, vol. 57, article 100755, 2025. doi: 10.1016/j.cosrev.2025.100755.
5. G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile Edge Intelligence for Large

- Language Models: A Contemporary Survey," IEEE Communications Surveys & Tutorials, vol. 27, no. 6, pp. 3820–3860, 2025. doi: 10.1109/COMST.2025.3527641.
6. Y. Zheng, Y. Chen, B. Qian, X. Shi, Y. Shu, and J. Chen, "A Review on Edge Large Language Models: Design, Execution, and Applications," ACM Computing Surveys, vol. 57, no. 8, article 209, 2025. doi: 10.1145/3719664.
 7. K. Sun, X. Wang, X. Miao, and Q. Zhao, "A Review of AI Edge Devices and Lightweight CNN and LLM Deployment," Neurocomputing, vol. 614, article 128791, 2025. doi: 10.1016/j.neucom.2024.128791.
 8. M. Navardi, R. Aalishah, Y. Fu, Y. Lin, H. Li, Y. Chen, and T. Mohsenin, "GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices," Proceedings of the AAAI Spring Symposium Series, vol. 5, no. 1, pp. 180–187, 2025. doi: 10.1609/aaais.v5i1.35586.
 9. S. Laskaridis, S. I. Venieris, A. Kouris, R. Li, and N. D. Lane, "The Future of Consumer Edge-AI Computing," arXiv:2210.10514, 2022. [Online]. Available: <https://arxiv.org/abs/2210.10514>.
 10. F. De Vita, D. Bruneo, and S. K. Das, "On-device Artificial Intelligence Solutions with Applications to Smart Environments," Future Generation Computer Systems, vol. 180, article 108373, 2026. doi: 10.1016/j.future.2026.108373.