1. Software Engineer & Masters Student, National School of Advanced Engineering, Polytech Yaoundé

ABSTRACT:

Background and Purpose: Edge Artificial Intelligence (AI) has emerged as a crucial solution for minimizing power consumption during real-time data processing in computing devices. Given the increasing demand for energy-efficient AI systems, this study systematically reviews Edge AI methods focused on optimizing energy efficiency. The research highlights existing challenges and explores future research directions to enhance Edge AI capabilities.

Methods: This study employs a systematic review methodology, analyzing recent advancements in Edge AI, including new hardware systems, model optimization techniques, and software development tools. The review considers the interplay between model complexity, hardware constraints, security concerns, and the trade-off between performance and energy efficiency.

Findings: The analysis reveals that optimizing Edge AI requires a multifaceted approach involving hardware innovations, lightweight models, and adaptive algorithms. Key challenges include balancing computational power with energy constraints, ensuring data security in edge environments, and maintaining real-time processing capabilities.

Theoretical Contributions: This research identifies three significant research avenues: (1) integrating neuromorphic computing to enhance efficiency and mimic biological neural processes, (2) leveraging federated learning to improve privacy-preserving model training across distributed edge devices, and (3) developing adaptive AI architectures that dynamically adjust computational resources based on workload demands.

Conclusions and Policy Implementations: To advance Edge AI, policymakers and industry leaders must prioritize the development of energy-efficient hardware, encourage research into low-power AI models, and establish regulatory frameworks for secure edge computing. Future work should explore interdisciplinary collaborations to further optimize Edge AI performance.

Keywords: Edge AI, Energy Efficiency, Low-Power Computing, Model Optimization, Neuromorphic Computing, Federated Learning

INTRODUCTION:

Edge Artificial Intelligence (AI) stands as a fundamental technology in modern computing by giving users the ability to process data instantly with decreased energy usage and speed-related limitations. Traditional AI systems depend on central server infrastructure for cloud computing because these servers analyze extensive data to create intelligent results. Implementing this method produces exceptional computational capability but it causes power consumption problems and network dependency problems as well as data privacy issues. Internet of Things devices along with autonomous systems and smart infrastructure have created an expanding market demand for localized AI processing which functions efficiently using constrained resources (1). Edge AI delivers solutions to industry needs through direct processing of AI models on edge devices which reduces dependence on cloud servers yet generates better energy performance. There are specific hurdles in edge AI deployment because of power restrictions and restricted processing capability together with model optimization requirements.

The adoption of Edge AI technology happens primarily because organizations require computing solutions that consume less energy. The traditional cloud-based AI utilizes extensive server centers for AI processing while edge devices that include smartphones and wearables and drones and autonomous vehicles run from restricted power supplies (2). AI models for battery-operated devices must operate efficiently while minimizing power usage because these systems demand low power consumption. Deep learning models designed for high-performance computing environments prove too intensive for edge deployment in their current state. They need modifications to work effectively on such systems. Numerous techniques developed by researchers and engineers have improved the way energy efficiency functions when running AI inference at the edge. The methods for enhancing edge device AI inference efficiency mainly consist of three categories: hardware optimizations and model compression techniques along with software-based solutions (3).

The development of hardware creates essential improvements which lead to greater Edge AI energy efficiency. Specialized low-power AI processors which include Google's Edge TPU and NVIDIA Jetson and ARM Cortex-M series fulfill their purpose of executing AI operations effectively in edge devices. The NPU technology built into these processors enhances AI speed processing while requiring minimal power consumption. The emerging neuromorphic computing framework based on human brain characteristics brings forth notable power efficiency benefits to the field (4). The Loihi from Intel and TrueNorth from IBM course information through event-based asynchronous computing techniques which enables them to reduce power requirements beyond traditional von Neumann systems. The combination of Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) has demonstrated potential as energy-efficient hardware components that enable optimized power utilization when performing customized AI accelerations. Such specialized hardware solutions enable Edge AI to perform high-speed inferences with minimal power requirements.

Model compression methods represent an essential component in decreasing the computational requirements of edge-deployed AI models. Quantization stands as one of the most popular computational techniques in neural networks to lower the sensitivity of numerical operations (5). Standard deep learning models handle their computations using 32-bit floats but reducing precision to 16-bit or 8-bit integer values cuts down power usage without compromising performance. Neural network efficiency increases through the pruning method that eliminates repetitive tiny parameters. The removal of unneeded weights in pruned models leads to diminished computational needs and therefore faster performance along with lower energy requirements (7). The knowledge distillation process enables the creation of efficient student models which duplicate the performance of larger teacher models. With this method edge devices can access deep learning knowledge systems without creating excessive expense in computing resources. Edge devices achieve efficient AI model performance while preserving accuracy levels through established compression practices (8).

Software-based techniques help achieve energy-efficient Edge AI by improving system-level AI execution procedures. The implementation of customized AI frameworks which excel in edge computing environments serves as an effective approach for energy-efficient operations. Several AI frameworks such as TensorFlow Lite, PyTorch

Mobile and ONNX Runtime give developers optimized AI model deployments that conserve system memory and truncate processing time. The developed frameworks achieve efficient performance on limited power systems without losing compatibility with traditional AI development methodologies (9). A critical software optimization method uses dynamic resource allocation since AI models transform their complexity definitions according to actual time resource constraints. The running speed and accurate delivery of AI models operating on mobile devices automatically adjusts according to battery strength and available processing capabilities. Federated learning presents a significant technique to execute AI model training across multiple distributed devices without relocating data (10). By using this framework organizations can maintain data security while simultaneously cutting down power usage during transmissions which benefits the operation of power-efficient AI applications.

The deployment of AI at the edge faces ongoing difficulties because of various built-in obstacles. The main obstacle in edge AI development appears in selecting between model sophistication levels and energy consumption requirements. Efficiency compared to size matters when determining model power consumption but smaller models often deliver inferior accuracy results than power-intensive larger ones. The search for optimal efficiency-performance means Edge AI experts and developers face a primary development challenge. The minimal processing power and memory capacity of edge devices makes it hard to execute deep learning applications which would otherwise reduce system speed and performance. Tools must operate compatibly on different edge devices that possess distinct processing engines due to hardware architecture variations which leads to getting optimally prepared AI models for diverse edge devices deployments.

Edge AI deployment faces major obstacles because of security along with privacy considerations. The decentralized nature of edge devices along with their limited resources makes them exposed to adversarial attacks and data breaches in addition to tampering events (11). The security measures for edge AI models differ from cloud-based systems since protection of edge models demands powerful on-device protection systems. Research seeks to tackle security issues through multiple strategies that include encrypted model execution and secure enclave solutions and adversarial defense systems. Security measures need further development to be deployed without causing detrimental effects on energy consumption.

This research aimed to review existing techniques that improve Edge AI system energy efficiency along with determining problems that arise during edge AI model implementation within resource-limited circumstances. Multiple issues persist in the development of energy-efficient AI processing at edge locations even though substantial improvements have been achieved. Reducing energy consumption creates an accuracy-performance trade-off since energy-saving measures usually result in performance degradation. Standardized AI solutions face development challenges because different edge devices possess heterogeneous nature which makes seamless operation across various hardware architectures difficult. The future success of Edge AI depends heavily on interdisciplinary work between AI technology and low-power computing and advanced security protocols to solve current technical barriers for real-time efficient computing operations.

2. Methods

2.1 Methodology

The review complies with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards so it can perform thorough assessments of appropriate research studies. The review analyzes energy-efficient computing methods used in Edge AI through the evaluation of improved hardware components and model optimization procedures together with software development frameworks. The methodology pursues stages that begin with literature search followed by screening and eligibility assessment to arrive at final inclusion of research materials. Scientists performed thorough database research through the combination of IEEE Xplore and ACM Digital Library with Springer and ScienceDirect along with Google Scholar. Boolean logical operators were used to develop the search terms leading to the retrieval of appropriate research results. The search utilized sets of keywords including ("Edge AI" OR "Edge Intelligence") AND ("Energy-Efficient Computing" OR "Low-Power AI") and ("Model Optimization" OR "Quantization" OR "Pruning") AND ("Hardware Acceleration" OR "AI Chips") as well as

("Federated Learning" OR "Neuromorphic Computing") and ("IoT" OR "Embedded Systems"). The overall database search produced 81 records but subsequent screening operations followed.

2.2 Study Selection and Screening

A complete selection process was implemented to identify studies. For the selected databases 81 records were retrieved at first but screening removed the duplicate records (n=16) before evaluation. The 65 research elements underwent further screening procedure by assessing their titles and abstract content for suitability to the research field. The application resulted in discarding 20 publications because they lacked direct relation to Edge AI combined with energy-efficient computing. The screening process yielded 45 full-text articles which were evaluated for eligibility yet seven papers were excluded because they lacked enough data or were non-peer-reviewed or insufficient in methodological rigor. Related search efforts resulted in 38 accepted studies for inclusion in the systematic review that delivered vital information about energy-efficient Edge AI strategies as presented in Figure 1.

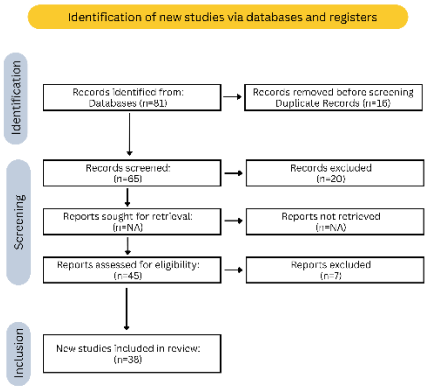

Figure 1: Prisma Flowchart

2.3 Inclusion and Exclusion Criteria

To maintain relevance and quality, the following selection criteria were applied:

- **Inclusion Criteria:**
  1. Peer-reviewed journal articles and conference proceedings.
  2. Research explicitly addressing energy efficiency in Edge AI.
  3. Studies proposing or evaluating model optimization techniques, hardware efficiency improvements, or software frameworks.
  4. Experimental studies reporting power consumption metrics, inference latency, and accuracy trade-offs.
- **Exclusion Criteria:**
  1. Papers focusing solely on cloud AI without an edge computing perspective.
  2. General AI studies without explicit discussion on energy efficiency.
  3. Non-English studies (due to translation constraints).
  4. Duplicate or redundant studies.

2.4 Data Extraction and Analysis

The selection process revealed appropriate studies which were grouped into hardware innovations, model optimization techniques and software frameworks sections. The research on hardware components investigated reduced power consumption for AI processors and neuromorphic systems as well as FPGA/ASIC-based acceleration techniques. The assessment of model optimization provided details about quantization methods and pruning operations and knowledge distillation techniques that could reduce computational requirements without sacrificing model accuracy. The analysis of software frameworks involved an assessment of Edge AI libraries as well as adaptive learning strategies and federated learning models. Power usage, system speed during inference and model accuracy together with trade-offs

between energy efficiency and throughput measurements were collected from every research study. Different methods underwent comparative assessment for evaluating their capacity to improve energy efficiency levels. The research reviewed case studies alongside real-world applications to both explain theoretical findings and find practical execution difficulties of Edge AI for low-power computing devices.

Results and Discussion

In this systematic review, several strategies were identified for improving energy efficiency in Edge AI systems, with a focus on hardware innovations, model optimization techniques, and software frameworks. These strategies aim to enhance computational efficiency while minimizing power consumption, a key challenge for Edge AI applications.

A significant portion of the research points to the development of hardware innovations to drive energy-efficient computations at the edge. Specialized low-power AI processors, such as Google's Edge TPU, NVIDIA Jetson, and ARM's Cortex-M series, have been designed specifically to reduce the power consumption of AI models running on edge devices. These processors are built to handle AI tasks with minimal energy expenditure while maintaining high processing speeds, which is critical for applications that demand real-time processing. Another key hardware advancement is neuromorphic computing, where processors like Intel's Loihi and IBM's TrueNorth emulate neural networks to achieve brain-inspired, low-power computations (12). These neuromorphic processors simulate the behavior of biological neurons to increase computational efficiency while significantly reducing power usage. In addition, energy-aware hardware components like Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) have been shown to offer tailored solutions for energy-efficient AI inference, especially in environments with strict power consumption requirements as shown in Table 1.

Table 1: Key Hardware Innovations for Energy-Efficient Edge AI

Hardware Innovation	Description	Example Devices/Technologies
Low-Power AI Processors	Specialized processors for energy-efficient AI computation	Google Edge TPU, NVIDIA Jetson, ARM Cortex-M series
Neuromorphic Computing	Brain-inspired processors that simulate neural networks	Intel Loihi, IBM TrueNorth
Energy-Aware FPGA and ASIC Designs	Customizable hardware solutions for efficient AI inference	Xilinx FPGAs, Custom ASICs

Alongside hardware developments, a number of model optimization techniques have emerged to reduce the computational demands of AI models while preserving performance. Quantization, for example, involves reducing the precision of model parameters from 32-bit to lower-bit representations, such as 8-bit, to reduce the amount of computation needed during inference. This reduction in model size leads to lower power consumption without significant degradation in performance. Similarly, pruning techniques remove redundant or non-essential parameters from neural networks, further optimizing them for energy efficiency (13). Another popular approach is knowledge distillation, which transfers the knowledge of larger, more complex models into smaller, more energy-efficient models as described in Table 2. By training a compact model to approximate the performance of a larger one, this technique reduces both computational overhead and energy usage.

Table 2: Model Optimization Techniques for Energy Efficiency

Technique	Description	Impact on Energy Efficiency
Quantization	Reduces the bit-width of model parameters, e.g., 32-bit to 8-bit	Lower computational load, reduced power consumption
Pruning	Removes redundant or non-essential parameters from the model	Reduces model size, optimizing energy usage
Knowledge Distillation	Transfers knowledge from a large model to a smaller, efficient model	Reduced computational overhead, energy savings

The review also highlighted the role of software frameworks in enhancing energy efficiency at the edge. Several

lightweight AI libraries, such as TensorFlow Lite, PyTorch Mobile, and ONNX Runtime, are specifically designed to optimize AI models for edge devices. These frameworks allow for efficient execution of models by reducing the computational burden while maintaining essential functionalities. Moreover, dynamic resource allocation techniques, where AI models adjust their complexity based on available energy resources, were found to be effective in ensuring that energy consumption is kept to a minimum while still delivering real-time results. These frameworks enable edge devices to operate more efficiently by adapting to varying conditions and workload demands.

The recent developments in Edge AI energy efficiency create positive opportunities but multiple implementation barriers persist. The main obstacle is that edge devices operate with restricted computational capabilities. The restricted features of these devices encompass limited memory and processing power and storage capacity thus limiting the allowable sophistication of deployed AI models. The energy efficiency challenge faces deep learning systems because many techniques designed to save energy often decrease model accuracy levels. The ongoing research focuses on achieving a proper balance between precise model performance and energy-saving capabilities in edge systems. Security concerns emerge when deploying AI models on edge devices because the hardware systems demonstrate heightened vulnerability to adversarial attacks together with privacy breaches. The management of sensitive data alongside maintaining model integrity is considered a major challenge for edge devices because they handle sensitive information. AI models require specialized optimization because edge devices present a combination of heterogeneous hardware that differs in computational performance specifications. The wide range of devices requires specialized optimization methods for models which makes the creation of energy-efficient systems more difficult as illustrated in Figure 2.

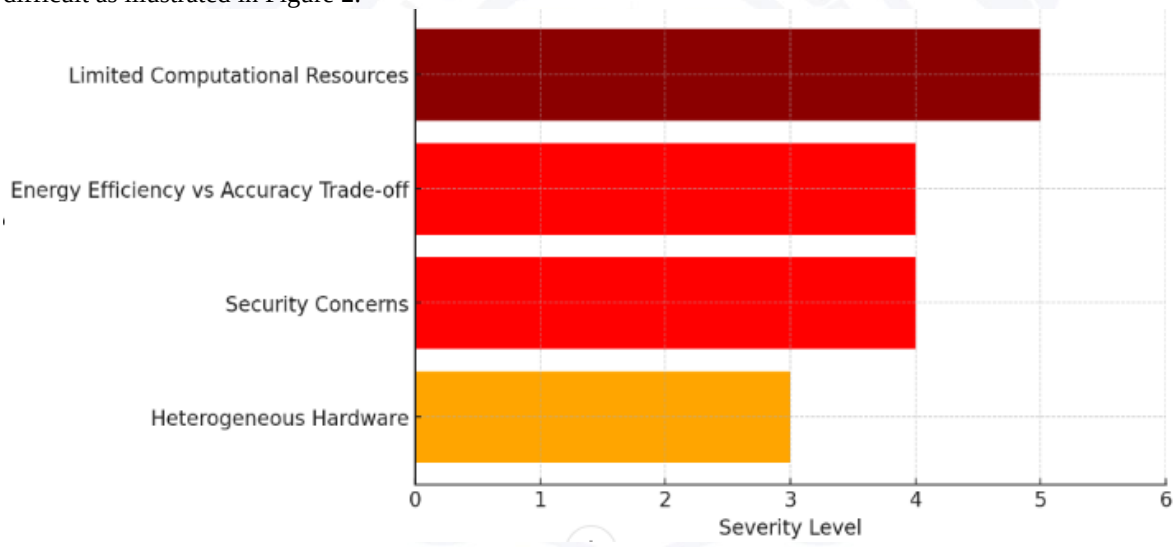

Figure 2: Challenges in Energy Efficient Edge AI

The future of energy-efficient Edge AI holds several promising directions. The decentralized training method known as federated learning enables various edge devices to develop models together by working on data separately without exchanging raw information. The method simultaneously preserves privacy and cuts down on data exchange with the cloud thus lowering energy usage. The next approach involves creating AI systems which autonomously determine their operational requirements using real-time power and computational resource capacities. The implementation of adaptive AI systems would lead to optimized performance when dealing with different edge devices while improving their energy-efficiency. Neuromorphic computing represents an important solution for low-power artificial intelligence systems because it duplicates brain neural networking functions. Research progress in both neuromorphic processor development and algorithm optimization will allow scientists to establish additional energy-efficient solutions. Edge-cloud hybrid AI systems that couple edge device processing capabilities to cloud computational

scalability would offer users a middle ground between system running speed and power consumption efficiency. The systems offload tasks to the cloud only at strategic times yet enable edge devices to complete complicated operations whenever necessary for maintaining maximum power efficiency.

Edge AI brings promising energy-efficient computing possibilities through the implementation of specific hardware units combined with optimized model deployment and optimized software frameworks. The implementation of Edge AI has shown impressive growth yet continuing research requires resolution of the issues regarding restricted processing power and balancing precision with efficiency and security matters. The research focus should move toward improving federated learning along with neuromorphic computing capabilities as well as adaptive AI architectures and hybrid edge-cloud AI systems to advance Edge AI efficiency and scalability.

Discussion

Edge Artificial Intelligence (AI) develops as an essential power-saving computational strategy which enables immediate data processing using decreased power utilization. The reviewed strategies reveal essential knowledge about hardware components and model optimization approaches and software frameworks needed to achieve edge computing power efficiency improvements. The energy performance of Edge AI systems has improved notably but several essential problems exist which require future research to achieve increased advancements in this domain.

Edge AI systems face performance restrictions because they have restricted access to computational capabilities on their deployed devices. The memory capacity along with processing power and storage capabilities that cloud-based systems provide are restricted on edge devices. The device constraints affect the upper limit of AI model complexity which can be utilized for deployment. Low-power processors manufactured by Google through their Edge TPU and NVIDIA with their Jetson products together with neuromorphic computing systems made by Intel using Loihi and IBM with TrueNorth help overcome hardware restrictions through efficient processing while maintaining high performance levels (14). The ongoing development of advanced processors creates new challenges when trying to achieve smooth deployment in various resource-limited environments. sensor and hardware variations across smartphones and industrial IoT sensors require numerous performance optimizations during model deployment and affect system total efficiency.

Energy efficiency stands in direct competition with accuracy according to identified research challenges. A significant number of model optimization techniques reviewed including quantization and pruning in addition to knowledge distillation seek to lower model size characteristics and computational needs (15). The energy-saving techniques result in diminished model performance yet these methods require sacrificing accuracy or operational precision. Real-time applications face a crucial trade-off because minor declines in accuracy lead to impacts on decision-making specifically within critical fields comprising healthcare alongside autonomous vehicles. The Edge AI field needs to solve fundamental research questions about finding balance between energy efficiency and performance metrics. Future research must concentrate on creating self-adaptive algorithms which modify programming sophistication throughout the span of real-time boundaries to preserve essential model accuracy.

Edge AI systems require immediate resolution of security concerns as one of their primary challenges. Although remote distributed edge devices handle sensitive information they remain open to adversarial attacks and privacy breaches (16). The process of securing energy-efficient AI systems proves to be complex due to several energy-saving methods like model compression and quantization that could compromise system security. Future AI model development priorities include building energy-efficient secure systems which incorporate protected hardware systems and end-to-end encrypted data processing and defensive techniques to secure lightweight models.

Federated learning introduces a pathway to boost Edge AI energy efficiency without compromising user data privacy levels. Through this decentralized method several edge devices can join forces to train models without needing raw data movement between the devices and cloud servers. The implementation of federated learning decreases both the need for data sharing and the resulting energy consumption from constant cloud-server connections. Scaling federated learning faces challenges in terms of creating efficient methods for model aggregation together with communication synchronization between devices having different processing capabilities. The study of improved federated learning

methods can enhance distributed Edge AI system performance by making them more scalable and energy efficient (17).

Neuromorphic computing demonstrates substantial abilities for creating ultra-low-power AI systems along with federated learning methods. The processing capability of neuromorphic processors that duplicates human brain operations allows major energy reductions against von Neumann architecture systems (18). Brain-inspired systems currently exist at their developmental infancy thus substantial work needs to be done in hardware creation and algorithm creation and real system implementation. At present neuromorphic computing represents a favorable method for reaching energy-efficient edge AI applications whereby continuous low-power processing is needed.

The hybrid edge-cloud AI framework proves itself as a successful method for achieving optimal energy efficiency. Through this dynamic model edge devices transfer computational workloads to the cloud as needed which preserves their energy efficiency while they execute complex operations. Utilizing this method provides edge devices with a suitable balance between their processing power and energy usage which permits advanced computation tasks to be transferred to the cloud system while maintaining live edge operations. Future studies need to examine hybrid edge-cloud designs to discover methods which will maintain performance quality and reduce power usage across different edge equipment in real-time applications.

While Edge AI presents significant opportunities for energy-efficient computing, several challenges must be addressed to fully realize its potential. These include optimizing model performance while minimizing energy usage, enhancing security, and addressing hardware constraints. With continued advancements in hardware, model optimization, and software frameworks, as well as promising research avenues such as federated learning, neuromorphic computing, and hybrid edge-cloud AI, the future of energy-efficient Edge AI holds great promise for a wide range of applications.

Conclusion

Edge AI offers a promising solution for energy-efficient computing by enabling real-time data processing on local devices with minimal power consumption. This systematic review highlights key strategies for optimizing energy efficiency, such as low-power processors, neuromorphic computing, and model optimization techniques like quantization and pruning. While significant progress has been made, challenges such as limited computational resources, trade-offs between accuracy and efficiency, and security concerns remain.

Looking ahead, promising research directions, including federated learning, neuromorphic computing, and hybrid edge-cloud AI, hold potential for further improving energy efficiency while addressing these challenges. As Edge AI continues to evolve, it is poised to drive sustainable computing solutions across diverse applications, from healthcare to autonomous systems.

References

1. Alam S., Yakopcic C., Wu Q., Barnell M., Khan S., & Taha T. Survey of deep learning accelerators for edge and emerging computing. *Electronics*. 2024;13(15):2988. https://doi.org/10.3390/electronics13152988
2. Amirsoleimani A., Alibart F., Yon V., Xu J., Pazhouhandeh M., Ecoffey S., et al. In-memory vector-matrix multiplication in monolithic complementary metal–oxide–semiconductor-memristor integrated circuits: design choices, challenges, and perspectives. *Advanced Intelligent Systems*. 2020;2(11). https://doi.org/10.1002/aisy.202000115
3. Bhardwaj K., Suda N., & Mărculescu R. Edgeal: a vision for deep learning in the IoT era. *IEEE Design and Test*. 2021;38(4):37-43. https://doi.org/10.1109/mdat.2019.2952350
4. Chang L., Zhu Z., Zhu Z., Yang S., Li W., & Zhou J. Energy-efficient spin-orbit torque MRAM operations for neural network processor. 2021. https://doi.org/10.1109/iscas51556.2021.9401449
5. Esser S., Merolla P., Arthur J., Cassidy A., Appuswamy R., Andreopoulos A., et al. Convolutional networks for fast, energy-efficient neuromorphic computing. *Proceedings of the National Academy of Sciences*. 2016;113(41):11441-11446. https://doi.org/10.1073/pnas.1604850113

6. Hao C., Dotzel J., Xiong J., Benini L., Zhang Z., & Chen D. Enabling design methodologies and future trends for edge AI: specialization and codesign. *IEEE Design and Test*. 2021;38(4):7-26. https://doi.org/10.1109/mdat.2021.3069952
7. Li W., Zhi-yuan H., Shen J., Luo D., Gao B., & Xie J. Distributed AI embedded cluster for real-time video analysis systems with edge computing. *MATEC Web of Conferences*. 2022;355:03036. https://doi.org/10.1051/matecconf/202235503036
8. Lu A., Lee J., Kim T., Karim M., Park R., Simka H., et al. High-speed emerging memories for AI hardware accelerators. *Nat Rev Electr Eng*. 2024;1(1):24-34. https://doi.org/10.1038/s44287-023-00002-9
9. Massarotto M., Saggini S., Loghi M., & Esseni D. Adiabatic leaky integrate-and-fire neurons with tunable refractory period in 180nm CMOS technology for ultra-low energy brain-inspired neuromorphic computing. 2024. https://doi.org/10.21203/rs.3.rs-4349574/v1
10. Muramatsu S., Nishida K., Ando K., & Asai T. Stochastic memory device based on a bistable system model with a simple analog circuit. *Nonlinear Theory and Its Applications, IEICE*. 2024;15(2):249-261. https://doi.org/10.1587/nolta.15.249
11. Muratore G., Rincon J., Julián V., Carrascosa C., Greco G., & Fortino G. Towards a dynamic edge AI framework applied to autonomous driving cars. 2020:406-415. https://doi.org/10.1007/978-3-030-51999-5_34
12. Oh J., Kim S., Choi J., Cha J., Im S., Jang B., et al. Memristor-based security primitives robust to malicious attacks for highly secure neuromorphic systems. *Advanced Intelligent Systems*. 2022;4(11). https://doi.org/10.1002/aisy.202200177
13. Sangarsu R. Securing the organization's sensitive data in the AI era. *Design of Single Chip Microcomputer Control System for Stepping Motor*. 2022:1-3. https://doi.org/10.47363/jaicc/2022(1)175
14. Satyasree E. Edge AI for real-time video analytics in surveillance systems. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2023;11(10):2269-2275. https://doi.org/10.17762/ijritcc.v11i10.8947
15. Schuman C., Patton R., Kulkarni S., Parsa M., Stahl C., Haas N., et al. Evolutionary vs imitation learning for neuromorphic control at the edge. *Neuromorphic Computing and Engineering*. 2022;2(1):014002. https://doi.org/10.1088/2634-4386/ac45e7
16. Sipola T., Alatalo J., Kokkonen T., & Rantonen M. Artificial intelligence in the IoT era: a review of edge AI hardware and software. 2022:320-331. https://doi.org/10.23919/fruct54823.2022.9770931
17. Véstias M., Duarte R., Sousa J., & Neto H. Moving deep learning to the edge. *Algorithms*. 2020;13(5):125. https://doi.org/10.3390/a13050125
18. Wulfert L., Kühnel J., Krupp L., Viga J., Wiede C., Gembaczka P., et al. AIFES: A next-generation edge AI framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024;46(6):4519-4533. https://doi.org/10.1109/tpami.2024.3355495