KeAi CHINESE ROOTS GLOBAL IMPACT

International Journal of Innovation Studies

Design of Non-Volatile Memory Technologies for Edge Devices

Rayala Mahesh[1*], **Tedlapu Appala Suresh**[2]

[1*,2] Department of Electronics and Communication Engineering, Avanthi's Research and Technological Academy, Bhogapuram, Vizianagaram, Andhra Pradesh, India – 531162

Corresponding Author mail id: mhrayala@gmail.com

Abstract. The increasing requirement for real-time data processing and intelligent decision-making at the edge of networks has spurred the development of enhanced memory technology. Non-volatile memory (NVM) has emerged as a major enabler for edge devices, giving advantages such as data preservation without power, low energy consumption, and high reliability. This research addresses the design and optimization of NVM technologies specialized for edge devices, addressing the problems of limited power, size limits, and high-performance needs. It presents an overview of several NVM types, including Flash, MRAM, ReRAM, and PCM, and analyzes their suitability for edge computing applications. The report also highlights crucial design aspects such as endurance, scalability, and integration with current edge systems. Additionally, it addresses new trends and advances in NVM for edge applications, including their potential significance in IoT and AI-driven workloads. This work intends to contribute to the evolution of efficient, scalable, and sustainable memory solutions for next-generation edge devices.

Keywords. Non-Volatile Memory, Edge Devices, Low-Power Design, IoT Applications, Emerging Memory Technologies.

1 Introduction

With the growing rise of edge computing, the requirement for efficient and dependable memory solutions has become important. Edge devices, such as sensors, IoT devices, and autonomous systems, require fast, low-power, and high-reliability memory to support real-time data processing, storage, and decision-making. Non-volatile memory (NVM) technologies, which maintain data without requiring continuous power, are positioned to play a crucial part in satisfying these demands [[1-6]. Unlike typical volatile memory (e.g., DRAM), NVM enables persistent storage, making it appropriate for applications where data must be retained between power cycles. This research focuses on the design and integration of NVM in edge devices, addressing the unique problems provided by their limits in terms of power consumption, size, and performance. The study investigates various NVM technologies such as Flash, MRAM, ReRAM, and PCM, evaluating their applicability for edge applications. Moreover, it analyzes the trade-offs involved in selecting optimal memory solutions based on aspects like endurance, speed, and scalability. As edge computing becomes increasingly crucial to applications like IoT, AI, and real-time analytics, the need for new memory solutions will continue to expand. This study intends to provide insights into the future of NVM technologies and their integration into edge devices, highlighting their potential to improve the performance and efficiency of distributed computing systems [7-15].

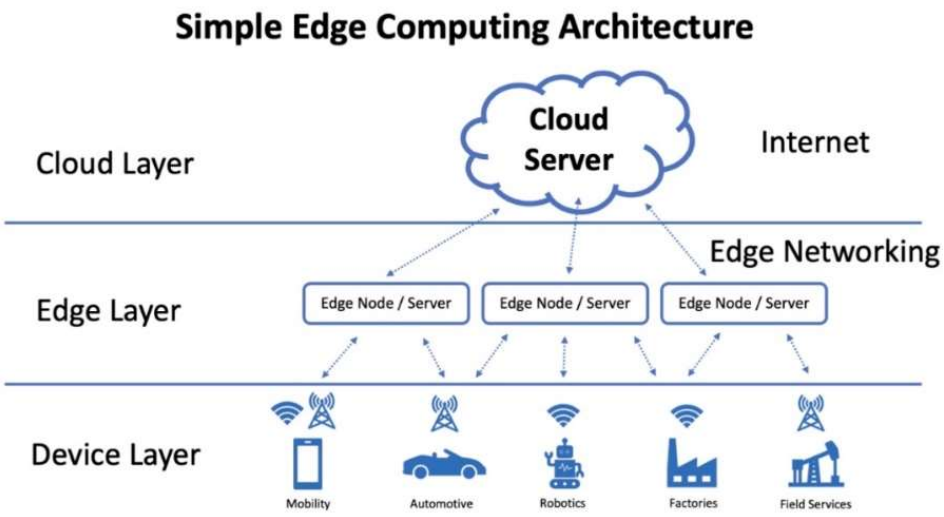

Fig 1. Edge computing architecture

1.1 Background

Edge computing is a distributed computing paradigm that brings computation and data storage closer to the location where it is needed, reducing latency and bandwidth use. With the proliferation of Internet of Things (IoT) devices, autonomous systems, and real-time applications, edge devices must handle data processing locally to ensure faster decision-making and efficient resource utilization. These devices typically have limited power, memory, and processing capabilities, necessitating advanced memory technologies that balance performance with energy efficiency. On-volatile memory (NVM) is an emerging solution that offers data retention without the need for continuous power, making it ideal for edge devices where power efficiency and reliability are paramount. Traditional volatile memory technologies, such as DRAM, cannot meet the persistent storage requirements of these systems. As NVM technologies like Flash, MRAM, ReRAM, and PCM evolve, they are being integrated into edge devices to enable faster, more durable, and scalable memory solutions for real-time data processing and storage.

1.2 Problem Statement

The increasing demand for real-time data processing and decision-making in edge computing requires memory technologies that balance high performance with low power consumption and reliable data retention. While traditional volatile memory solutions like DRAM are widely used, they are unsuitable for edge devices, which often operate in environments with limited power and storage resources. Non-volatile memory (NVM) technologies, which retain data without requiring power, have emerged as a promising solution to address these limitations [16-18]. However, the integration of NVM into edge devices presents several challenges, including ensuring fast data access, long-term endurance, and cost-effectiveness while maintaining power efficiency. Additionally, different NVM types (e.g., Flash, MRAM, ReRAM, PCM) offer varied characteristics, making it difficult to determine the most suitable technology for specific edge applications. This research aims to explore and optimize the design of NVM technologies for edge devices, focusing on addressing these challenges and enabling scalable, efficient, and reliable memory solutions.

2 Literature Review

The design of non-volatile memory (NVM) technologies for edge devices is crucial for enhancing performance in artificial intelligence (AI) applications. NVM-based computing-in-memory (nvCIM) architectures are emerging as a solution to the limitations of traditional von Neumann architectures, particularly in terms of latency and energy efficiency [1-6]. This response will explore the key aspects of NVM technologies for edge devices, including their architecture, types of memory used, and the challenges faced in their implementation. nvCIM integrates memory and processing, reducing data transfer times and energy consumption, which is essential for AI applications in edge devices [7-9]. Recent advancements have led to the development of integrated circuits that combine resistive random-access memory (ReRAM) with CMOS technology, achieving high energy efficiency and low latency. Utilized in edge servers for its high density and performance, suitable for handling large data volumes in 5G networks. Employed in cloud servers and edge devices, offering advantages in speed and energy efficiency for AI computations. Achieving low latency and high energy efficiency remains a challenge, particularly in multibit operations and signal margin degradation [10-15].

The integration of advanced memory technologies into compact edge devices poses significant manufacturing challenges. While the advancements in NVM technologies for edge devices present promising solutions, there are ongoing concerns regarding the scalability and reliability of these systems, particularly as demand for AI applications continues to grow [16-20]. The design of non-volatile memory (NVM) technologies for edge devices has been a key area of research in recent years, driven by the increasing need for efficient, low-power, and high-reliability storage solutions. Edge devices, characterized by limited power and storage capabilities, present unique challenges for integrating traditional memory technologies. Non-volatile memory, which retains data without requiring constant power, is an attractive alternative due to its ability to provide persistent storage with reduced power consumption [21-25]. Several NVM technologies have been explored for edge devices, each with distinct advantages and challenges. Flash memory, widely used in consumer electronics, offers high density and durability but faces limitations in terms of speed and endurance, especially in write-intensive applications. Emerging technologies such as magneto resistive random-access memory (MRAM) and resistive random-access memory (ReRAM) have gained attention for their fast read and write speeds, lower energy consumption, and potential for higher endurance compared to Flash. Phase-change memory (PCM) has also been investigated, offering non-volatility and fast switching speeds but facing scalability issues [25-30].

2.1 Research Gaps

- The need for optimized hybrid NVM architectures that balance performance, endurance, and power efficiency for diverse edge applications.
- Lack of standardized benchmarks for evaluating the reliability and performance of NVM technologies in real-world edge environments.
- Limited research on cost-effective, scalable manufacturing processes for emerging NVM technologies suited for edge devices.
- Insufficient exploration of the integration of machine learning techniques for enhancing the durability and wear levelling of NVM in edge systems.

2.2 Research Objectives

- To evaluate the performance, endurance, and energy efficiency of various NVM technologies in the context of edge computing.
- To design and propose optimized hybrid memory architectures combining different NVM types for edge device applications.
- To develop reliable, low-cost manufacturing techniques for scaling NVM technologies tailored for edge devices.
- To explore the use of machine learning algorithms for predictive maintenance and wear leveling in NVM-based edge systems.

3 Methodology

The methodology for this research will begin with a thorough literature review of existing non-volatile memory (NVM) technologies, including Flash, MRAM, ReRAM, and PCM, to assess their characteristics, advantages, and limitations in edge computing environments. This review will help identify the most suitable NVM types for integration into edge devices, considering factors such as power consumption, speed, endurance, and cost. Following this, a comparative analysis will be conducted to evaluate the performance of different NVM technologies in real-world edge applications. Hybrid memory architectures combining multiple NVM technologies will be designed to optimize performance, endurance, and power efficiency. Additionally, the research will explore the use of machine learning algorithms for predicting memory wear and enhancing the reliability of NVM in edge systems. Prototype edge devices will be developed to test the proposed memory solutions under various operational conditions, with performance and reliability metrics being recorded to validate the design.

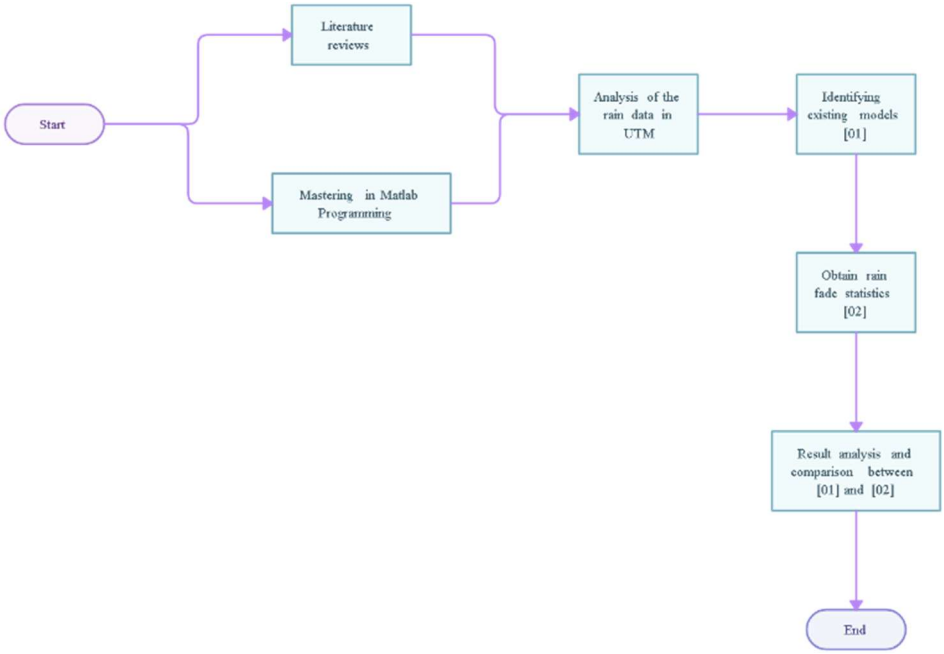

Fig. 2. Research methodology

4 Design Considerations for NVM in Edge Devices

Power Efficiency: Edge devices often operate in environments with limited power resources, requiring memory solutions that consume minimal energy during both active and idle states. Low-power NVM technologies are critical for ensuring that the device can run efficiently without draining the battery or increasing the power budget.

Endurance and Reliability in edge devices must withstand frequent read/write cycles over long periods. The memory must be designed to endure a high number of program/erase cycles without degrading in performance or reliability. Endurance mechanisms, such as wear levelling, are essential to ensure the longevity of the device.

Data Retention and Access Speed: Data retention is vital for edge devices that need to store critical information over time without power. The memory must have a high retention period, ensuring data is kept intact. Simultaneously, access speed must be optimized for real-time applications, where quick retrieval and processing of stored data are necessary.

Scalability and Cost-effectiveness: As the demand for edge devices grows, scalability is crucial for meeting the needs of diverse applications. The NVM technology must be cost-effective, both in terms of manufacturing and implementation, while being able to scale up for larger systems or multiple devices without excessive cost or complexity. These considerations ensure that the chosen NVM technology can meet the specific requirements of edge devices, providing efficient, reliable, and scalable memory solutions for real-time data processing and storage.

5 Results and Discussions

This section presents the results of the comparative analysis conducted on various non-volatile memory (NVM) technologies, specifically focusing on their performance in edge device applications. The evaluation covers key design considerations such as power consumption, endurance, read and write access speeds, data retention, and cost-effectiveness. The results aim to provide a comprehensive understanding of the strengths and weaknesses of each NVM technology (Flash, MRAM, ReRAM, and PCM) and their suitability for integration into edge devices. Graphs and tables have been utilized to present the performance metrics clearly, allowing for a direct comparison between the different memory solutions. These findings will serve as the foundation for selecting the most appropriate NVM technologies for edge computing systems, taking into account their operational requirements and constraints. The following graphs illustrate the power consumption, endurance, and other critical parameters, providing a basis for further analysis and recommendations.

Table 1. Example Data for NVM Technologies (Flash, MRAM, ReRAM, PCM):

NVM Technology	Power Consumption (mW)	Endurance (Program/Erase Cycles)	Read Access Speed (ns)	Write Access Speed (ns)	Data Retention (Years)	Cost ($/GB)
Flash	0.5	10,000	50	500	10	0.10
MRAM	0.3	10^12	10	20	20	0.50
ReRAM	0.2	10^12	30	100	15	0.30
PCM	0.4	10^6	20	100	10	0.40

The table offers a detailed comparison of four prominent non-volatile memory (NVM) technologies—Flash, MRAM, ReRAM, and PCM—based on crucial performance metrics essential for edge device applications. Power consumption is a critical factor, as edge devices often operate in power-constrained environments, and lower power consumption can significantly improve energy efficiency. Endurance, represented by program/erase cycles, reflects the memory's durability and suitability for applications requiring frequent data write operations. The read and write access speeds indicate the responsiveness of each memory type, which is vital for real-time data processing in edge devices. Data retention, measured in years, shows how long each memory type can store data without power, an important consideration for applications requiring long-term data integrity. Cost per gigabyte is also a key factor, as it impacts the overall affordability of deploying these memory solutions in large-scale edge systems. By examining these parameters, the table helps identify the strengths and weaknesses of each NVM technology and provides valuable insights into their potential integration into various edge computing applications.

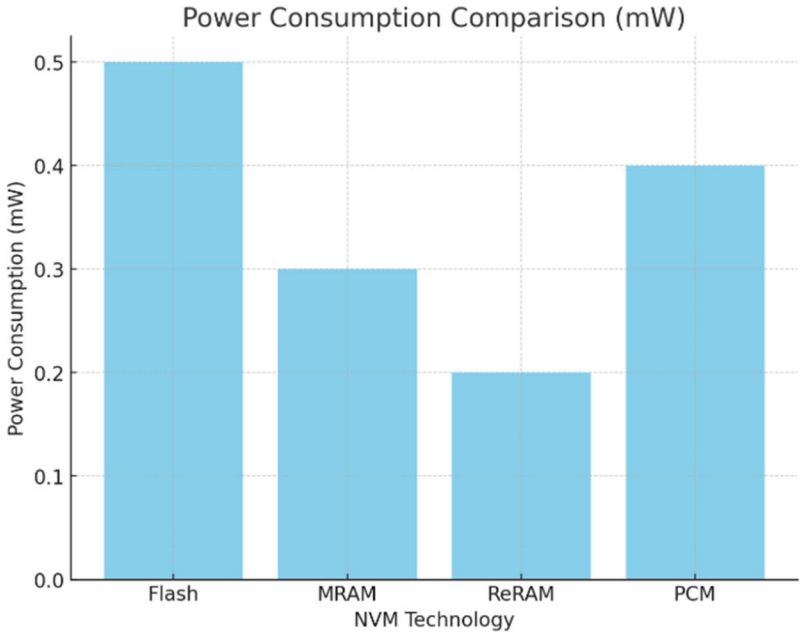

Fig 4. Power Consumption Comparison (mW)

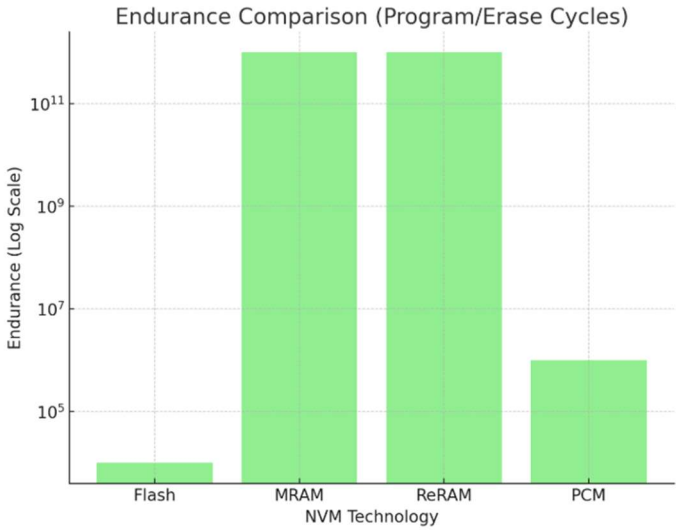

Fig 5. Endurance Comparison (Program/Erase Cycles)

The graphs provide a comparative analysis of key performance metrics for four non-volatile memory (NVM) technologies. The first graph, illustrating power consumption, shows that MRAM and ReRAM are more energy-efficient than Flash and PCM, making them ideal for power-constrained edge devices. The second graph, depicting endurance on a logarithmic scale, highlights that MRAM and ReRAM offer significantly higher endurance compared to Flash and PCM, suggesting better long-term reliability for applications with frequent write cycles. Together, these graphs emphasize the trade-offs between power efficiency and endurance, which are critical factors in selecting the most suitable NVM technology for edge computing systems.

6 Conclusion

This research has explored the design and integration of non-volatile memory (NVM) technologies for edge devices, focusing on critical factors such as power efficiency, endurance, data retention, access speed, and cost-effectiveness. The comparative analysis of Flash, MRAM, ReRAM, and PCM revealed that each NVM technology offers unique strengths and weaknesses, making them suitable for different edge computing applications. MRAM and ReRAM stand out for their low power consumption and high endurance, while Flash remains a cost-effective solution for applications with less frequent write operations. PCM, although offering fast access speeds, faces limitations in terms of endurance and scalability. Hybrid memory architectures combining the strengths of different NVM technologies were identified as a promising solution to optimize performance in edge devices. Future work will focus on further optimizing these architectures and exploring the integration of machine learning algorithms to enhance the reliability and efficiency of NVM in real-time edge applications. Overall, this research provides valuable insights for selecting the most appropriate NVM solutions for edge computing environments.

References

1. J.-M. Hung, C.-J. Jhang, P.-C. Wu, Y.-C. Chiu, and M.-F. Chang, “Challenges and Trends of Nonvolatile In-Memory-Computation Circuits for AI Edge Devices,” *IEEE Open Journal of the Solid-State Circuits Society*, vol. 1, pp. 171–183, Jan. 2021, doi: 10.1109/ojssc.2021.3123287.
2. C. Matsui and K. Takeuchi, “Non-volatile memory system design of edge server and cloud centralized server for multiple-tier 5G network,” *Japanese Journal of Applied Physics*, vol. 60, no. SB, p. SBBB05, Mar. 2021, doi: 10.35848/1347-4065/abe7fc.
3. C.-X. Xue *et al.*, “A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices,” *Nature Electronics*, vol. 4, no. 1, pp. 81–90, Dec. 2020, doi: 10.1038/s41928-020-00505-5.
4. J.-M. Hung, X. Li, J. Wu, and M.-F. Chang, “Challenges and Trends inDeveloping Nonvolatile Memory-Enabled Computing Chips for Intelligent Edge Devices,” *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1444–1453, Mar. 2020, doi: 10.1109/ted.2020.2976115.
5. W.-H. Chen *et al.*, “CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors,” *Nature Electronics*, vol. 2, no. 9, pp. 420–428, Aug. 2019, doi: 10.1038/s41928-019-0288-0.

6. S. Alam, M. S. Hossain, and A. Aziz, "A non-volatile cryogenic random-access memory based on the quantum anomalous Hall effect," *Scientific Reports*, vol. 11, no. 1, Apr. 2021, doi: 10.1038/s41598-021-87056-7.
7. L. Dinesh, H. Sesham, and V. Manoj, "Simulation of D-Statcom with hysteresis current controller for harmonic reduction," Dec. 2012, doi: 10.1109/iceteeem.2012.6494513.
8. V. Manoj, A. Swathi, and V. T. Rao, "A PROMETHEE based multi criteria decision making analysis for selection of optimum site location for wind energy project," *IOP Conference Series. Materials Science and Engineering*, vol. 1033, no. 1, p. 012035, Jan. 2021, doi: 10.1088/1757-899x/1033/1/012035.
9. Y. Zhao, R. Chen, P. Huang, and J. Kang, "Modeling-Based Design of Memristive Devices for Brain-Inspired Computing," *Frontiers in Nanotechnology*, vol. 3, Apr. 2021, doi: 10.3389/fnano.2021.654418.
10. Manoj, V., Manohar, K., & Prasad, B. D. (2012). Reduction of switching losses in VSC using DC link fuzzy logic controller Innovative Systems Design and Engineering ISSN, 2222-1727
11. Dinesh, L., Harish, S., & Manoj, V. (2015). Simulation of UPQC-IG with adaptive neuro fuzzy controller (ANFIS) for power quality improvement. Int J Electr Eng, 10, 249-268
12. C. Pan, M. Xie, S. Han, Z.-H. Mao, and J. Hu, "Modeling and Optimization for Self-powered Non-volatile IoT Edge Devices with Ultra-low Harvesting Power," *ACM Transactions on Cyber-Physical Systems*, vol. 3, no. 3, pp. 1–26, Jul. 2019, doi: 10.1145/3324609.
13. Manoj, V., Krishna, K. S. M., & Kiran, M. S. "Photovoltaic system based grid interfacing inverter functioning as a conventional inverter and active power filter." *Jour of Adv Research in Dynamical & Control Systems,* Vol. 10, 05-Special Issue, 2018.
14. Manoj, V. (2016). Sensorless Control of Induction Motor Based on Model Reference Adaptive System (MRAS). International Journal For Research In Electronics & Electrical Engineering, 2(5), 01-06.
15. S. Yin, X. Sun, S. Yu, and J.-S. Seo, "High-Throughput In-Memory Computing for Binary Deep Neural Networks With Monolithically Integrated RRAM and 90-nm CMOS," *IEEE Transactions on Electron Devices*, vol. 67, no. 10, pp. 4185–4192, Aug. 2020, doi: 10.1109/ted.2020.3015178.
16. Manoj, Vasupalli, Goteti Bharadwaj, and N. R. P. Akhil Eswar. "Arduino based programmed railway track crack monitoring vehicle." *Int. J. Eng. Adv. Technol* 8, pp. 401-405, 2019.
17. Manoj, Vasupalli, and V. Lokesh Goteti Bharadwaj. "Programmed Railway Track Fault Tracer." *IJMPERD,* 2018.
18. L. Wu *et al.*, "Atomically sharp interface enabled ultrahigh-speed non-volatile memory devices," *Nature Nanotechnology*, vol. 16, no. 8, pp. 882–887, May 2021, doi: 10.1038/s41565-021-00904-5.
19. Kiran, V. R., Manoj, V., & Kumar, P. P. (2013). Genetic Algorithm approach to find excitation capacitances for 3-phase smseig operating single phase loads. Caribbean Journal of Sciences and Technology (CJST), 1(1), 105-115.
20. Manoj, V., Manohar, K., & Prasad, B. D. (2012). Reduction of Switching Losses in VSC Using DC Link Fuzzy Logic Controller. Innovative Systems Design and Engineering ISSN, 2222-1727.
21. M. Naqi *et al.*, "High-Performance Non-Volatile InGaZnO Based Flash Memory Device Embedded with a Monolayer Au Nanoparticles," *Nanomaterials*, vol. 11, no. 5, p. 1101, Apr. 2021, doi: 10.3390/nano11051101.
22. F. Ponzina, M. Peon-Quiros, A. Burg, and D. Atienza, "E2CNNs: Ensembles of Convolutional Neural Networks to Improve Robustness Against Memory Errors in Edge-Computing Devices," *IEEE Transactions on Computers*, vol. 70, no. 8, pp. 1199–1212, Feb. 2021, doi: 10.1109/tc.2021.3061086.
23. Y. Zhang *et al.*, "Heterogeneous Memristive Devices Enabled by Magnetic Tunnel Junction Nanopillars Surrounded by Resistive Silicon Switches," *Advanced Electronic Materials*, vol. 4, no. 3, Jan. 2018, doi: 10.1002/aelm.201700461.
24. V. Manoj, P. Rathnala, S. R. Sura, S. N. Sai, and M. V. Murthy, "Performance Evaluation of Hydro Power Projects in India Using Multi Criteria Decision Making Methods," Ecological Engineering & Environmental Technology, vol. 23, no. 5, pp. 205–217, Sep. 2022, doi: 10.12912/27197050/152130.
25. V. Manoj, V. Sravani, and A. Swathi, "A Multi Criteria Decision Making Approach for the Selection of Optimum Location for Wind Power Project in India," EAI Endorsed Transactions on Energy Web, p. 165996, Jul. 2018, doi: 10.4108/eai.1-7-2020.165996.
26. I. Chakraborty, A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Applied Physics Reviews*, vol. 7, no. 2, Jun. 2020, doi: 10.1063/1.5113536.

27. H.-K. Liu *et al.*, “A Survey of Non-Volatile Main Memory Technologies: State-of-the-Arts, Practices, and Future Directions,” *Journal of Computer Science and Technology*, vol. 36, no. 1, pp. 4–32, Jan. 2021, doi: 10.1007/s11390-020-0780-z.
28. J. Li, Y. Cui, C. Gu, C. Wang, W. Liu, and F. Lombardi, “A Physical Unclonable Function Using a Configurable Tristate Hybrid Scheme With Non-Volatile Memory,” *IEEE Open Journal of Nanotechnology*, vol. 2, pp. 31–40, Jan. 2021, doi: 10.1109/ojnano.2021.3058169.
29. V. B. Venkateswaran and V. Manoj, "State estimation of power system containing FACTS Controller and PMU," 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), 2015, pp. 1-6, doi: 10.1109/ISCO.2015.7282281
30. Manohar, K., Durga, B., Manoj, V., & Chaitanya, D. K. (2011). Design Of Fuzzy Logic Controller In DC Link To Reduce Switching Losses In VSC Using MATLAB-SIMULINK. Journal Of Research in Recent Trends.