A Study on Memristor-Based CNN Models for Scalable Pattern Detection

Prince Patel[1]

Assistant Professor, Department of Artificial Intelligence & Data Science, Sankalchand Patel University, Visnagar, India [1]

pcpatelai_spce@spu.ac.in

Abstract: The MNIST and Fashion MNIST datasets are used in this work to create a Convolutional Neural Network (CNN) model specifically designed for picture classification applications. CNNs have attracted a lot of interest because they are good at identifying complex patterns in visual data. Our research attempts to classify handwritten numbers and fashion items with great accuracy by utilizing these capabilities. With around 98% accuracy on the MNIST dataset and about 90% accuracy on the Fashion MNIST dataset, the model exhibits impressive performance after being painstakingly trained and verified. Using Python's PyTorch framework, the architecture is developed, utilizing deep learning tasks' flexibility and efficiency. We improve the network parameters through rigorous testing to maintain computational efficiency while maximizing classification accuracy. Our results highlight the CNN's promise for practical uses in object recognition and classification as well as its resilience when processing a variety of picture datasets. Furthermore, our findings demonstrate how effective CNN is in classifying images compared to more conventional approaches, presenting encouraging directions for future computer vision research and advancement. By highlighting the CNN's capacity to meet the expanding needs of AI applications, this work opens the door for developments in automated image analysis, object detection, and pattern recognition, among other subjects. CNNs become increasingly important as computing systems advance because they can handle the difficulties presented by large, varied datasets and a variety of application domains.

Keywords: Memristor, Dion-Jacobson Hybrid Perovskite, Flexible, Resistive Switching, Hebbian Learning, Synaptic Behavior, Convolutional Neural Network, Artificial Neural Networks, Feed forward Neural Networks, Recurrent Neural Networks, Pattern Recognition.

I. INTRODUCTION

In the realm of computing, the traditional von Neumann architecture, while effective for solving complex mathematical problems, encounters limitations with big data, the Internet of Things (IoT), and artificial intelligence (AI). The "Von Neumann bottleneck" occurs due to the separation of memory and processor, leading to high energy usage and slow processing speeds. In contrast, the human brain, with its neural networks of around 10^{11} neurons and 10^{15} synapses, excels at parallel processing with minimal energy consumption. Neuromorphic computing, inspired by the brain's architecture, aims to overcome these limitations [1]. Artificial Neural Networks (ANN) has evolved through three generations: perceptron modelling, deep neural networks (DNN), and spiking neural networks (SNNs). SNNs, the third generation, operate with improved speed and energy efficiency compared to DNNs by using spiking neurons and linked synapses. ANN can be categorized into feed forward neural networks (FFNN) and recurrent neural networks (RNN), where RNN is better suited for dynamic data, while FFNN processes static data. The convergence of neuromorphic computing and ANN shows promise in areas like pattern recognition and complex sensing. Researchers face challenges in developing and enhancing computing architectures to meet the demands of contemporary applications. Despite the enduring use of the Von Neumann architecture, issues like power consumption, latency, and scalability persist, driving the need for new approaches. This has led to increased research and development of neuromorphic hardware, particularly memristors, to usher in a new era of efficient and resilient computing technologies [2].

Memristors provide a unique chance to seamlessly combine computing and storage capabilities by mimicking the performance of biological synapses. Researchers are starting to investigate how they could transform neural network topologies, especially in the field of deep learning, by utilizing their innate features. Memristors' capacity for parallel processing is well suited to convolutional neural networks' (CNNs) needs, allowing them to perform tasks like feature extraction with greater efficiency than ever before. Furthermore, memristor arrays' applicability for CNN architectures is further highlighted by their implementation as convolutional kernels. Given its significant influence on personal performance and health, the importance of stress detection and diagnosis in today's culture cannot be overemphasized. When it comes to categorizing stress levels based on different physiological inputs, such as brain activity, traditional methods using CNNs have made great progress. But given the intrinsic drawbacks of traditional CNN architectures, it is important to investigate alternate strategies that can provide more accuracy and dependability.

The neuroimaging method known as functional near-infrared spectroscopy (fNIRS) shows promise in quantifying the reactions of hemoglobin to stress. fNIRS circumvents the limitations of other imaging modalities and offers portability and safety, making it a great option for practical applications. Particularly well-suited for stress detection is its exceptional temporal resolution, which allows it to quantify variations in both oxygenated and deoxygenated hemoglobin. Enhancing stress detection techniques through the integration of fNIRS data with memristor-based neural networks is an intriguing approach. Scientists want to improve the accuracy of classification between control and stress groups by converting memristor conductances into neural network weights. Memristor-based CNNs (M-CNNs) are an innovative method that utilizes the special qualities of memristors to potentially outperform conventional CNNs [3].

In this study, we provide an in-depth analysis of the use of fNIRS data to apply memristor-based neural networks for stress detection. Using memristor-style DenseNet (M-DenseNet) and M-CNNs, two cutting-edge neural network designs, we test their performance against traditional CNNs and DenseNet models using publicly accessible fNIRS-stress datasets. We show that memristor-based techniques are effective at increasing classification accuracy and reliability in stress detection tasks through thorough testing and analysis. Moreover, we tackle the difficulties brought about by non-ideal effects in memristor-based architectures, suggesting hardware-supportive implementations and training plans resistant to these effects. We offer important insights and recommendations for the useful integration of these technologies in real-world applications by methodically examining the feasibility of memristor-based neural networks and optimizing approaches at both the hardware and algorithm levels. This study aims to advance stress detection techniques by utilizing the capabilities of neural networks based on memristors combined with fNIRS data. Our goal is to further the development of neuromorphic computing paradigms and their applications in important fields like healthcare and performance monitoring by means of theoretical analysis and empirical validation.

II. BACKGROUND AND RELATED STUDIES

In the realm of deep learning research, Convolutional neural networks (CNNs), which have a basic design consisting of alternating convolution and max-pooling layers, frequently combined with a few fully connected layers, are the cornerstone of many studies in the field of deep learning research. In these layers, dropout techniques are often used for network regularization, and nonlinear activation functions are frequently used. Additionally, procedures for dropping connections are incorporated for additional regularization. Although the assessment of these Deep Convolutional Neural Network (DCNN) structures usually takes place on general-purpose computing platforms such as CPUs, GPGPUs, and multicore systems, optimization attempts encompass both computational and structural aspects. Aiming to reduce computational parameters and increase overall accuracy, structural optimization projects also cut down on electricity and computing time [4]. The optimization SqueezeNet stands out for its ability to minimize computational complexity and provide faster processing while maintaining power efficiency. Furthermore, new techniques like binary weight networks, which is used to demonstrate high-precision training with only binary weight values, have been made possible by advances in low-precision implementations. In recent times, ternary weight-based CNNs have surfaced, utilizing ternary connected networks in frameworks such as IBM Eedn. The argument over broad vs deep convolutional networks is still going strong, as evidenced by the numerous studies examining full-precision implementations and examining the trade-offs between depth and width in CNN architectures. The deep learning community's constant quest to optimise DCNN models for effectiveness, speed, and performance across a wide range of applications is highlighted by this complex environment [5].

Recent research titled "Do Deep Learning Need to Be Deep?" goes into great detail on how network architecture affects recognition accuracy. The research findings clearly indicate that deeper networks, which are defined by more layers that enable better feature embedding, perform consistently better than wider networks, which are defined by more neurons and larger feature maps inside each layer. Further studies examining the effect of network topology on accuracy with the same amount of factors confirm that deeper networks perform better than wider ones [6]. Extra study demonstrates that, when parameter counts are held constant, shallow networks are unable to achieve accuracy levels comparable to those of deep networks, which supports this empirical finding. The combined results support the idea that deeper networks perform better than shallower networks on a regular basis, indicating the importance of network depth in deep learning system optimization for better recognition performance and overall efficacy [7].

This inquiry concerns the suitability of the ternary connect technique for Deep Convolutional Neural Networks (DCNN) on IBM's TrueNorth platform. This claim may or may not be true; however, IBM's TrueNorth system has demonstrated potential when it comes to deep learning techniques that make use of binary weights (0, 1). This is an appropriate approach for integrating deep learning into this architecture. A new path for energy-efficient deep learning implementations on neuromorphic hardware was opened in 2016 by IBM with the release of the Eedn deep learning framework, which was designed specifically for TrueNorth. This framework shows remarkable accuracy, especially in picture classification tasks, and a noteworthy power efficiency. It is therefore necessary to assess the impact of network architectures on recognition performance for TrueNorth, in contrast to traditional CPU, GPGPU, and multicore system designs. Empirical evaluations of different DCNN architectures on a range of datasets provide information that can be utilized to guide the effective creation of DCNN models intended for the TrueNorth system. In addition to assisting in the improvement of currently available energy-efficient models, this project has the potential to generate new models with improved recognition performance in the field of neuromorphic computing.

III. METHODOLOGY

3.1 Neuro-Cognitive Synapse Network (NCSN)

Convolutional neural networks and brain-inspired architecture are combined in neuromorphic computing, which mimics the flexibility of biological synapses by utilizing memristors as key components. Memristors, which mimic natural synapses in terms of synaptic plasticity, allow artificial neurons to communicate effectively with one another. This architecture, which resembles the intricate synaptic network of the brain, encourages parallelism in information processing. This method departs from standard sequential computing in that it allows for simultaneous computations across numerous paths. A neuromorphic model like this signals the beginning of a new age in artificial intelligence by utilizing the brain's parallel and holistic processing abilities. Memristor integration holds the potential to improve neural network performance, efficiency, and energy economy, paving the way for the creation of AI systems that are almost as sophisticated as the human brain in terms of learning and adaptability.

The human brain's enormous network of neurons and synapses allows it to process information in parallel, which allows it to execute tasks like picture recognition and data classification better than traditional computers. Synapses alter joint strength or weight by allowing electrical signals to flow between neurons, as seen in Figure 1. The brain is incredibly complex, with 10,000 connections per neuron and about 100 billion neurons. Long-Term Depression (LTD) and Long-Term Potentiation (LTP) are two examples of mechanisms where synaptic weights are updated throughout the exchange of excitatory and inhibitory postsynaptic potentials between neurons. Learning and memory formation depend on this process, which is called synapse weight modification based on the interval between pre-synaptic and post-synaptic action potentials, or spikes (STDP). Because of their non-volatile properties, low power consumption, and small on-chip footprint, memristors, which can imitate LTD and LTP behaviors, offer appealing attributes for artificial synaptic devices. Memristors are attractive options for developing artificial intelligence systems because of these features.

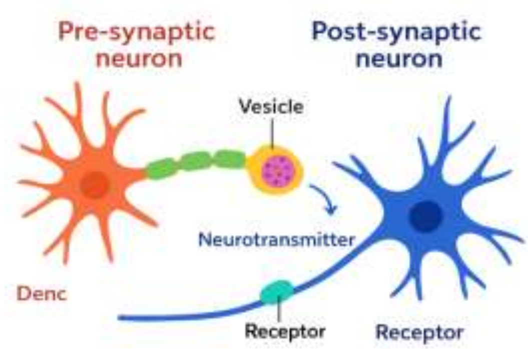

Fig. 1 synaptic junction is the connection between a pre-synaptic neuron and a post-synaptic neuron.

Memristive devices are subject to particular demands from neural network applications; these include low resistance fluctuations, low device-to-device variability, and a broad resistance range to support a range of resistance states. Elevated absolute resistance helps to minimize power dissipation, and durability is essential for reprogramming and training. On the other hand, resistance drift over time is a problem, impacting synapse weight and neural network accuracy. Enhancements in material engineering, system design, and circuit design are essential to overcome this. In material engineering, threading dislocations can improve programming control and switching uniformity. Circuit-level design can use modules of two series memristors in conjunction with the smallest transistor to encode resistance ratio and reduce resistance drift.Device variation can be reduced by system design using protocols

like write-verify functions in closed-loop peripheral circuits. Linear and symmetric weight updates are required for both Long-Term Depression (LTD) and Long-Term Potentiation (LTP) in neural network training. This is accomplished by means of optimized programming pulses that maintain memristor integrity, regardless of whether fixed-width or fixed-amplitude voltage pulses are employed. Unfortunately, there is an inevitable increase in energy consumption when using intricately programmed pulses to modify memristor resistance.

3.2 Dataset

This implementation uses popular benchmarks for digit and object recognition tasks, the MNIST and Fashion MNIST datasets. These datasets are well-known in the machine learning community for testing how well models perform on classification tasks, offering consistent standards for evaluating algorithmic improvements in image recognition.

3.2.1 MNIST

MNIST, a well-known benchmark for image classification, has 10,000 test examples and 60,000 training samples. Every picture is a 28x28 grayscale depiction of the numbers 0 through 9. We did not use any data augmentation techniques in our experiment; instead, we only resized the input samples as needed. The dataset's diversity is demonstrated by the provided photographs, which also act as the basis for assessing the effectiveness of categorization models.

Fig. 2 MNIST Dataset

3.2.2 Fashion MNIST

Zalando generated the Fashion-MNIST dataset, which consists of 10,000 test cases and 60,000 training examples of 28x28 pixel grayscale images. These pictures depict a range of fashion accessories and are divided into ten different categories. With the same image size and separation of training and testing sets, Zalando created Fashion-MNIST to replace the MNIST dataset. This dataset offers a standardized platform for assessing model performance and advances in the field of fashion detection and categorization, hence facilitating the benchmarking of machine learning algorithms in this area.

Fig. 3 Fashion MNIST

IV. RESULTS AND DISCUSSION

Excitatory postsynaptic currents (EPSC) are the subject of a supervised learning study that uses Figure 4 to illustrate the possibilities of flexible hybrid perovskite memristors in neuromorphic computing. The natural learning, memory, and forgetting mechanisms of the human brain must be replicated in order to mimic its functions. It is possible to simulate synaptic plasticity and gain insight into the neuromorphic uses of memristors by adjusting input pulses or stimuli in artificial synapses, which allows for the understanding of changes in synaptic strength. In our devices, the most significant reduction in negatively charged V_{Pb}' and V_{BDA}' back diffusion occurs when the second stimulation occurs before the first EPSC fully vanishes. This results in a more prominent buildup of V_{Pb}' and VBDA' at the $BDAPbI_4$-PCBM-Ag interface, which causes discernible conductance changes. Furthermore, as shown in Figure 4, the non-linearity factor was calculated by using non-identical voltage pulses with fixed widths and increasing amplitudes in potentiation and depression (P&D). The formula for this computation was $NL = \frac{\left[\frac{(Gmax-Gmin)}{2}+Gmin\right]-Gs}{\left[\frac{(Gmax-Gmin)}{2}+Gmin\right]} X\ 100\%$. The results show that for both depression and potentiation, the non-linearity factor for P&D employing non-identical pulses approaches around 0%.

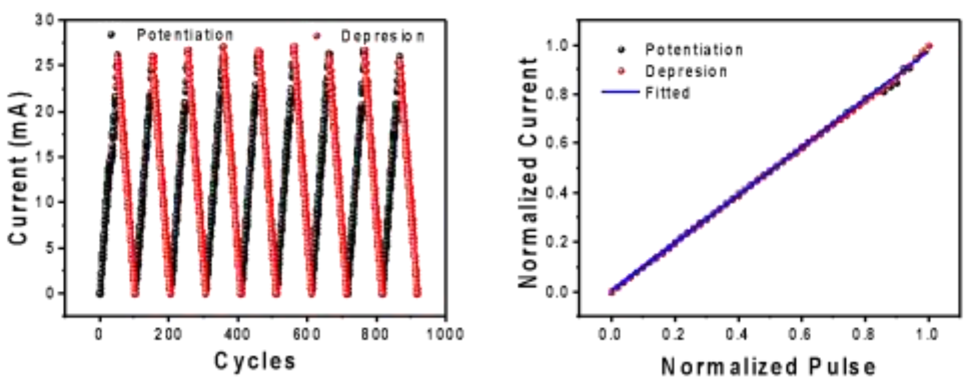

Fig. 4 Visualization of Normalized P&D Characteristics

We analyzed various neural network architectures with different layer and core configurations on the MNIST and Fashion MNIST datasets, observing performance fluctuations based on core count. Additionally, we explored how recognition accuracy changes with network structure, even when using the same core quantity.

4.1 MNIST

The number of feature maps and groups in the default implementation network was changed in order to assess the performance of the MNIST dataset. The results, which are shown in Figure 5, show that performance improved noticeably as the number of cycles rose. This improvement was especially visible in bigger networks. The testing accuracy improvement to roughly 99.07% with deeper networks containing more cores and a broader structure was another finding from Figure 5. As shown in Figure 5, further tests using a broader version of the network but keeping the same number of cycles showed that deeper networks consistently performed better than wider ones. Comparing deeper structures to their wider counterparts, this comparison demonstrates how superior deeper architectures are in improving testing accuracy when used in the MNIST dataset evaluation context.

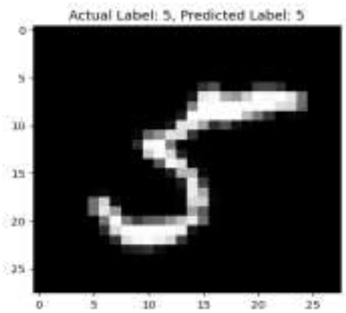

Fig. 5 Sample Test Image from the MNIST Dataset

4.2 Fashion MNIST

The feature maps and groups of the default implementation network were modified after analyzing the Fashion MNIST dataset's performance. The results, which are shown in Figure 6, showed that performance improved noticeably with an increase in cycles, especially in bigger networks. As Figure 6 illustrates, testing accuracy increased to about 94.07% for deeper networks with more cores and a wider topology. Further experiments using a larger network version while preserving the cycle count consistently preferred deeper networks over wider ones, demonstrating their better performance in improving testing accuracy in the context of evaluating the Fashion MNIST dataset. This comparison highlights the deeper design's superior performance over their wider equivalents, highlighting their importance in improving testing accuracy.

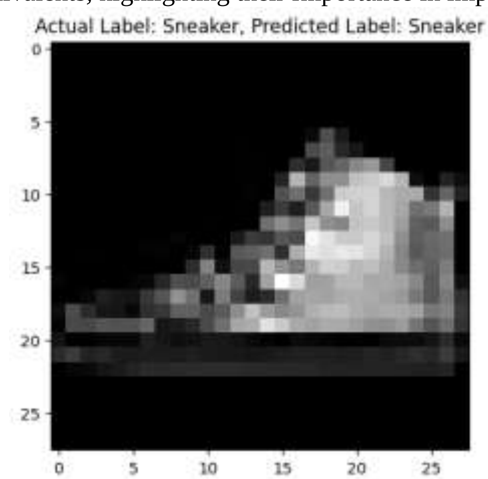

Fig. 6 Sample Test Image from the Fashion MNIST Dataset

V. CONCLUSION

Our study demonstrates the memristive application-specific Dion-Jacobson memristor devices' exceptional adaptability. These devices are made utilizing a specialized solution technique. We demonstrated impressive performance on pattern recognition tasks with the MNIST and Fashion MNIST datasets by utilizing depression and potentiation approaches, yielding test accuracies of roughly 98% and 91%, respectively. Considering that convolutional neural networks are naturally capable of guided learning, this accomplishment is quite noteworthy. We successfully illustrate Hebb's learning rule using depression and potentiation, highlighting its usefulness in various contexts. Additionally, we have further confirmed the effectiveness of our model through associative learning by using the MNIST and Fashion MNIST datasets for tasks like random number and picture prediction. These results suggest that synthetic synapses made of hybrid DJ perovskite have great potential for neuromorphic and flexible computing applications, as they mimic the complex functions of the human brain.

REFERENCES

[1] Q. Tian, et al., “Metal oxide-based resistive switching memristors for neuromorphic computing,” Journal of Materials Chemistry C, vol. 13, no. 24, pp. 12046–12065, 2025.

[2] S. Duan, et al., “Memristor-based cellular nonlinear/neural network: Design, analysis, and applications,” IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 6, pp. 1202–1213, 2014.

[3] M. Patel, et al., “Insights of $BDAPbI_4$-based flexible memristor for artificial synapses and in-memory computing,” ACS Omega, vol. 9, no. 47, pp. 46841–46850, 2024.

[4] P. Patel, et al., “Memristor-based neuromorphic computing and artificial neural networks for computer vision and AI—Applications,” in Biomedical Imaging: Advances in Artificial Intelligence and Machine Learning, Singapore: Springer Nature, 2024, pp. 307–322.

[5] P. Yao, et al., “Fully hardware-implemented memristor convolutional neural network,” Nature, vol. 577, no. 7792, pp. 641–646, 2020.

[6] S. Bak, et al., “Memristor-based CNNs for detecting stress using brain imaging signals,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 8, no. 1, pp. 140–149, 2023.

[7] G. Leng, et al., “Research on construction and simulation methods of memristor-based CNN,” in 2025 IEEE 8th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), vol. 8, IEEE, 2025.