International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

DESIGN AND IMPLEMENTATION OF AN ADC/DAC-FREE WALSH–HADAMARD TRANSFORM BASED NEURAL NETWORK ACCELERATOR USING BIT-PLANE PROCESSING FOR LOW-POWER EDGE INTELLIGENCE APPLICATIONS

SYEDA WAJEEHA DANISH[1], Dr. Y. RAGHAVENDER RAO[2]

[1]**PG Scholar,** VLSI SYSTEM DESIGN, ECE DEPARTMENT, JNTUH University College of Engineering, Sultanpur. wajeehadanish2000@gmail.com

[2]**Professor & Vice Principal,** ECE DEPARTMENT, JNTUH University College of Engineering, Sultanpur. yraghavenderrao@gmail.com

Abstract: The rapid growth of artificial intelligence and deep neural networks has created a significant demand for efficient hardware accelerators capable of delivering high computational performance under strict power and resource constraints. Conventional neural network accelerators rely heavily on multiplication-intensive operations, large memory bandwidth, and analog-to-digital (ADC) and digital-to-analog (DAC) conversion circuits, resulting in increased power consumption, silicon area, and processing latency. This paper presents a novel ADC/DAC-free neural network accelerator based on the Walsh–Hadamard Transform (WHT) and bit-plane processing techniques for low-power edge intelligence applications. The proposed architecture replaces conventional multiply-accumulate operations with addition and subtraction by exploiting the binary orthogonal properties of the Walsh–Hadamard matrix. A Block Walsh–Hadamard Transform (BWHT) compute core is developed to perform frequency-domain neural computation using fixed transform coefficients, thereby eliminating the need for trainable weight multiplications.

To achieve high-precision processing without analog conversion circuits, multibit input data is decomposed into individual bitplanes and processed sequentially from the most significant bit to the least significant bit. The architecture further incorporates comparator-based output generation, soft-thresholding for sparsity enhancement, and an early termination mechanism that dynamically halts unnecessary computations to reduce energy consumption and latency. The complete system is modeled and implemented using Verilog HDL and is optimized for FPGA realization. Experimental analysis demonstrates that the proposed accelerator significantly reduces computational complexity, memory overhead, power consumption, and hardware resource utilization while maintaining reliable neural computation.

The elimination of multipliers, ADCs, and DACs, combined with frequency-domain processing and bit-plane computation, makes the proposed architecture an efficient and scalable solution

for next-generation edge AI systems, embedded intelligence platforms, and real-time machine learning applications.

Keywords: Neural Network Accelerator, Walsh–Hadamard Transform, Block Walsh–Hadamard Transform, ADC/DAC-Free Architecture, Bit-Plane Processing, FPGA Implementation, Edge Intelligence, Frequency-Domain Computing, Low-Power Design, Hardware Acceleration, Verilog HDL, Soft Thresholding, Early Termination, Compute-in-Memory, Artificial Intelligence Hardware.

I. INTRODUCTION

Artificial Intelligence (AI) and Deep Neural Networks (DNNs) have transformed modern computing by enabling intelligent decision-making across diverse applications such as computer vision, healthcare, autonomous vehicles, speech recognition, robotics, and edge intelligence systems. Over the past decade, deep learning architectures including Convolutional Neural Networks (CNNs), Residual Networks (ResNet), and MobileNet have demonstrated remarkable success in extracting complex features and achieving state-of-the-art performance in various pattern recognition tasks. However, the growing complexity of these models has significantly increased computational requirements, creating major challenges for deployment on resource-constrained edge devices. Most deep neural networks rely heavily on multiply-accumulate (MAC) operations, which dominate both computational workload and energy consumption. As the depth of neural networks increases, the number of arithmetic operations grows rapidly, resulting in high latency, excessive power dissipation, and increased hardware complexity. These challenges become particularly critical in edge computing environments where devices operate under strict limitations of power, memory, and computational resources.

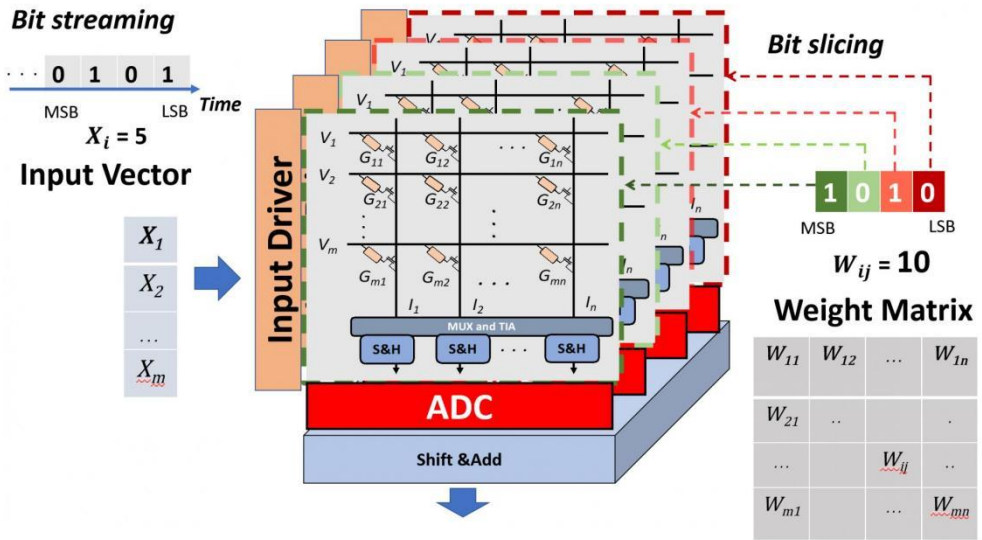

Fig 1: Bit-Streaming and Bit-Slicing Based Compute-in-Memory Crossbar Architecture with ADC and Shift-and-Add Reconstruction

Traditional neural network accelerators implemented on CPUs, GPUs, and Tensor Processing Units (TPUs) provide high computational throughput but consume considerable energy and require extensive

hardware resources. Although Field Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) offer improved efficiency, they still depend heavily on multiplication-intensive processing elements. Multipliers occupy significant silicon area and contribute substantially to overall power consumption. Furthermore, neural network computations require frequent access to large volumes of weights and feature maps stored in memory. The repeated transfer of data between memory and processing units leads to the well-known memory bottleneck problem, where data movement consumes more energy than the actual computation. As modern neural networks continue to scale in size and complexity, these limitations hinder their deployment in low-power embedded systems and real-time edge intelligence applications.

To address these challenges, researchers have explored various optimization techniques including network pruning, quantization, weight sharing, low-precision arithmetic, and approximate computing. While these approaches reduce computational complexity and memory requirements to some extent, they do not fundamentally eliminate the dependence on multiplication operations. Moreover, aggressive optimization often introduces trade-offs between computational efficiency and inference accuracy. Therefore, there is a growing need for alternative computational paradigms that can significantly reduce arithmetic complexity while maintaining reliable neural network performance. One promising direction is frequency-domain neural computation, where neural operations are transformed into a domain that enables more efficient mathematical processing.

Among various frequency-domain techniques, the Walsh–Hadamard Transform (WHT) has emerged as an attractive alternative due to its binary and orthogonal structure. Unlike conventional transforms such as the Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT), which require complex multiplications and floating-point arithmetic, the Walsh–Hadamard Transform operates using only addition and subtraction. The transform matrix consists exclusively of +1 and −1 coefficients, allowing multiplication operations to be replaced with simple sign changes and arithmetic additions. This property makes WHT highly suitable for hardware implementation and low-power digital processing. Furthermore, the Fast Walsh–Hadamard Transform (FWHT) algorithm reduces computational complexity through a recursive butterfly architecture that supports efficient parallel execution. These characteristics enable the development of neural accelerators that significantly reduce computational overhead while maintaining the ability to extract meaningful feature representations.

Recent research has demonstrated that transform-domain neural networks can effectively replace certain trainable convolution and projection layers with fixed orthogonal transformations. By utilizing predefined Walsh–Hadamard basis functions, neural computation can be reformulated as structured signal decomposition rather than conventional weight multiplication. This approach eliminates the need for storing and updating large numbers of trainable parameters, thereby

reducing memory requirements and simplifying hardware design. In addition, the orthogonality of the Walsh matrix ensures efficient feature representation and reduces redundancy in transformed data. Such transform-based architectures offer significant potential for low-power edge intelligence systems where computational efficiency is a primary design objective.

Another important challenge in modern neural accelerators is the dependence on analog-to-digital converters (ADCs) and digital-to-analog converters (DACs), particularly in compute-in-memory and mixed-signal architectures. Although analog computation offers improved energy efficiency and parallelism, ADCs and DACs introduce substantial overhead in terms of power consumption, silicon area, and processing latency. In many compute-in-memory implementations, data conversion circuits account for a significant portion of total system energy consumption. Moreover, high-resolution converters increase hardware complexity and limit scalability. Eliminating ADC and DAC components therefore represents an important step toward achieving highly efficient neural processing systems.

To overcome these limitations, the proposed work introduces an ADC/DAC-free neural network accelerator based on the Walsh–Hadamard Transform and bit-plane processing. Instead of relying on analog signal conversion, the architecture directly processes digital input data through a bit-plane decomposition mechanism. Multibit input vectors are separated into individual biplanes beginning with the most significant bit and progressing toward the least significant bit. Each bitplane is processed independently using a Block Walsh–Hadamard Transform compute core, enabling high-precision computation through simple binary operations. This strategy eliminates the need for complex analog circuitry while maintaining computational accuracy and scalability.

The proposed architecture further incorporates comparator-based output generation, which replaces conventional analog-to-digital conversion with lightweight digital decision-making. By comparing positive and negative accumulated partial sums, the system generates output bits efficiently without requiring expensive quantization hardware. In addition, a soft-thresholding mechanism is introduced to enforce sparsity in the transformed domain. This operation suppresses insignificant coefficients while preserving important signal components, thereby reducing computational load and improving overall efficiency. The architecture also includes an early termination mechanism that dynamically monitors partial computation results and halts unnecessary processing when subsequent bitplanes are unlikely to influence the final output. This feature significantly reduces switching activity, latency, and power consumption.

The complete accelerator is modeled and implemented using Verilog HDL and is optimized for FPGA realization. The proposed design combines the advantages of Walsh–Hadamard frequency-domain processing, bit-plane computation, comparator-based decision making, sparsity enhancement, and early termination into a unified hardware framework. By eliminating multipliers, trainable transformation weights, ADCs, and

DACs, the architecture achieves substantial reductions in computational complexity and resource utilization. Experimental evaluation demonstrates that the proposed accelerator provides a scalable, energy-efficient, and hardware-friendly solution for neural network inference. Consequently, the presented architecture represents a promising approach for next-generation edge intelligence systems, embedded artificial intelligence platforms, Internet of Things devices, wearable electronics, and real-time machine learning applications where low power consumption, reduced latency, and efficient hardware utilization are critical design requirements.

II. LITERATURE SURVEY

The rapid advancement of deep learning has resulted in significant improvements in image recognition, speech processing, healthcare diagnostics, and intelligent automation systems. However, the computational complexity associated with deep neural networks (DNNs) has become a major challenge for deployment on resource-constrained edge devices. Conventional neural network architectures such as AlexNet, VGGNet, ResNet, and MobileNet rely heavily on multiply-accumulate (MAC) operations, which account for the majority of computational workload and energy consumption. Krizhevsky et al. demonstrated the effectiveness of deep convolutional neural networks for large-scale image classification; however, the proposed architecture required extensive computational resources and high-performance hardware platforms. Subsequent architectures improved accuracy but further increased the number of trainable parameters and arithmetic operations, making efficient hardware implementation increasingly difficult.

To address these challenges, researchers have explored hardware acceleration techniques using Graphics Processing Units (GPUs), Tensor Processing Units (TPUs), Application-Specific Integrated Circuits (ASICs), and Field Programmable Gate Arrays (FPGAs). Jouppi et al. introduced the Tensor Processing Unit, a specialized hardware accelerator designed to improve the efficiency of neural network inference. Although TPUs achieved substantial performance gains, they still relied on multiplication-intensive computations and consumed significant power. FPGA-based accelerators have also been widely investigated because of their flexibility, parallel processing capabilities, and lower power consumption. Zhang et al. developed an FPGA-based convolutional neural network accelerator that improved throughput through parallel execution of convolution operations. However, the architecture continued to depend on MAC units, resulting in considerable hardware complexity and resource utilization.

To overcome the limitations associated with multiplication-intensive processing, researchers have explored frequency-domain neural computation techniques. Frequency-domain approaches transform input data into orthogonal basis representations where computations can be performed more efficiently. Traditional methods based on the Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) have been applied for signal compression and feature extraction. However, these transforms re-

quire complex multiplications and floating-point arithmetic, which limit their suitability for low-power hardware implementations. Consequently, alternative transforms with simpler computational structures have gained attention.

The Walsh–Hadamard Transform (WHT) has emerged as a highly attractive frequency-domain technique due to its binary orthogonal structure. Unlike Fourier-based transforms, the Walsh–Hadamard matrix contains only +1 and −1 coefficients, enabling multiplication-free computation using addition and subtraction operations. Fino and Algazi introduced efficient Fast Walsh–Hadamard Transform (FWHT) algorithms that significantly reduce computational complexity through recursive butterfly structures. Subsequent research demonstrated that WHT-based processing can be effectively applied to signal analysis, image compression, and machine learning applications. The transform's deterministic nature, orthogonality, and computational simplicity make it particularly suitable for hardware realization on FPGA and ASIC platforms.

Several researchers have investigated the integration of Walsh–Hadamard Transform into neural network architectures. Dao et al. proposed structured transform layers that replace conventional weight matrices with fixed orthogonal transforms, thereby reducing parameter count and computational complexity. Their results showed that transform-based neural layers could achieve competitive performance while requiring significantly fewer arithmetic operations. Similar studies demonstrated that Walsh–Hadamard projections can effectively replace channel mixing and linear transformation layers within deep neural networks. These approaches eliminate the need for storing large weight matrices and improve computational efficiency. However, most existing implementations focus primarily on algorithmic optimization and do not fully address hardware-level challenges such as data conversion overhead and energy consumption.

Low-precision neural computation has also received significant attention as a means of reducing hardware complexity. Binary Neural Networks (BNNs), Ternary Neural Networks (TNNs), and quantized neural networks restrict weights and activations to lower precision representations. Courbariaux et al. introduced BinaryConnect, which replaces multiplications with simple binary operations during training and inference. Similarly, Hubara et al. developed Binary Neural Networks capable of performing inference using binary arithmetic. Although these approaches reduce computational requirements, aggressive quantization often degrades classification accuracy and requires careful retraining procedures. Furthermore, many low-precision architectures still depend on multipliers and conversion circuits for intermediate computations.

Bit-plane processing has emerged as an alternative approach for achieving high-precision computation using simple binary operations. In bit-plane architectures, multibit input data is decomposed into individual bitplanes and processed sequentially. This approach enables scalable precision control while reducing datapath complexity. Several studies have demonstrated that bit-plane processing improves hardware utilization

and supports adaptive precision computation. Furthermore, bit-plane methods are highly compatible with transform-domain processing because each bitplane can be independently transformed using simple arithmetic operations. However, existing bit-plane accelerators often focus on arithmetic optimization without fully exploiting transform-domain sparsity and computation reduction mechanisms.

Recent research has explored ADC/DAC-free neural computation architectures to eliminate the overhead associated with analog signal conversion. Comparator-based processing techniques have been proposed as lightweight alternatives to traditional analog-to-digital conversion. These methods generate output decisions using simple comparison operations rather than full-resolution quantization. While comparator-based approaches reduce power consumption and hardware complexity, limited research has combined such techniques with Walsh–Hadamard Transform processing and bit-plane computation in a unified architecture. Furthermore, existing systems rarely incorporate sparsity-driven optimization techniques such as soft-thresholding and early termination to further improve efficiency.

From the literature review, it is evident that significant progress has been achieved in neural network compression, hardware acceleration, compute-in-memory architectures, transform-domain processing, and low-precision computation. However, most existing approaches address these challenges independently. Conventional neural accelerators continue to depend on multiplication-intensive operations, while analog compute-in-memory systems suffer from ADC/DAC overhead. Transform-based neural networks reduce computational complexity but often lack hardware-efficient implementation strategies.

Therefore, a research gap exists for a unified architecture that combines Walsh–Hadamard Transform-based frequency-domain processing, bit-plane computation, comparator-based output generation, soft-thresholding, and early termination in a completely ADC/DAC-free framework. The proposed work addresses this gap by developing a Verilog-based neural accelerator that eliminates multipliers and data converters while achieving efficient, scalable, and low-power neural computation suitable for edge intelligence applications.

III. PROPOSED METHODOLOGY

The proposed work introduces an ADC/DAC-free Walsh–Hadamard Transform (WHT) based neural network accelerator that performs frequency-domain neural computation using bit-plane processing techniques. The primary objective of the proposed architecture is to reduce computational complexity, power consumption, memory overhead, and hardware resource utilization while maintaining reliable neural network inference performance. Unlike conventional neural accelerators that depend on multiplication-intensive multiply-accumulate (MAC) operations and analog-to-digital (ADC) and digital-to-analog (DAC) conversion circuits, the proposed system utilizes a fixed Walsh–Hadamard transformation matrix and comparator-based output generation to achieve efficient hardware implementation.

The overall architecture consists of seven major processing stages: Input

Buffer, Bit-Plane Extraction, Block Walsh–Hadamard Transform (BWHT) Compute Core, Signed Partial Sum Unit, Comparator Unit, Soft Thresholding Unit, and Early Termination Controller. Initially, the multibit input vector is stored in the input buffer, which provides synchronized data transfer and ensures stable operation of the processing pipeline. The buffered input data is subsequently passed to the bit-plane extraction module, where each input sample is decomposed into individual bitplanes beginning from the Most Significant Bit (MSB) and progressing toward the Least Significant Bit (LSB). This decomposition enables the system to process high-precision input data using simple binary arithmetic operations while maintaining overall numerical accuracy.

The extracted bitplanes are supplied to the Block Walsh–Hadamard Transform compute core. The BWHT module performs frequency-domain transformation using a predefined orthogonal Walsh matrix consisting exclusively of +1 and −1 coefficients. Because the transform matrix contains only binary sign values, conventional multiplication operations are completely eliminated and replaced with simple addition and subtraction operations. The BWHT core utilizes a Fast Walsh–Hadamard Transform (FWHT) butterfly architecture that efficiently computes transform coefficients through recursive processing stages. This significantly reduces computational complexity compared with traditional convolution and matrix multiplication operations used in deep neural networks.

Following the transformation stage, the generated coefficients are forwarded to the Signed Partial Sum (PSUM) Unit. This module separates positive and negative transformed values into independent accumulation paths. The positive and negative partial sums are continuously updated as each bitplane is processed. By maintaining separate accumulation channels, the architecture supports efficient comparator-based decision making and eliminates the need for high-resolution analog computation circuits. The accumulated partial sums represent the transformed neural features in the Walsh frequency domain and provide the foundation for subsequent classification and decision operations.

The Comparator Unit replaces the conventional Analog-to-Digital Converter typically found in compute-in-memory and mixed-signal neural accelerators. Instead of quantizing analog values using complex conversion circuitry, the comparator directly evaluates the relationship between positive and negative accumulated sums. If the positive accumulation exceeds the negative accumulation, a logical output of one is generated; otherwise, a logical output of zero is produced. This approach significantly reduces hardware complexity, power consumption, and processing latency while maintaining functional correctness. The comparator output corresponding to each processed bitplane is forwarded to the Output Bit Accumulator.

The Output Bit Accumulator reconstructs the final multibit output by combining comparator outputs from all processed bitplanes. Each generated output bit is assigned a weight corresponding to its bit position, enabling accurate reconstruction of high-precision transformed values. This process allows the architecture to achieve scalable precision without

requiring wide arithmetic data paths or high-complexity processing elements. Consequently, the system supports flexible precision control and can be adapted to various edge computing applications with different performance and energy requirements.

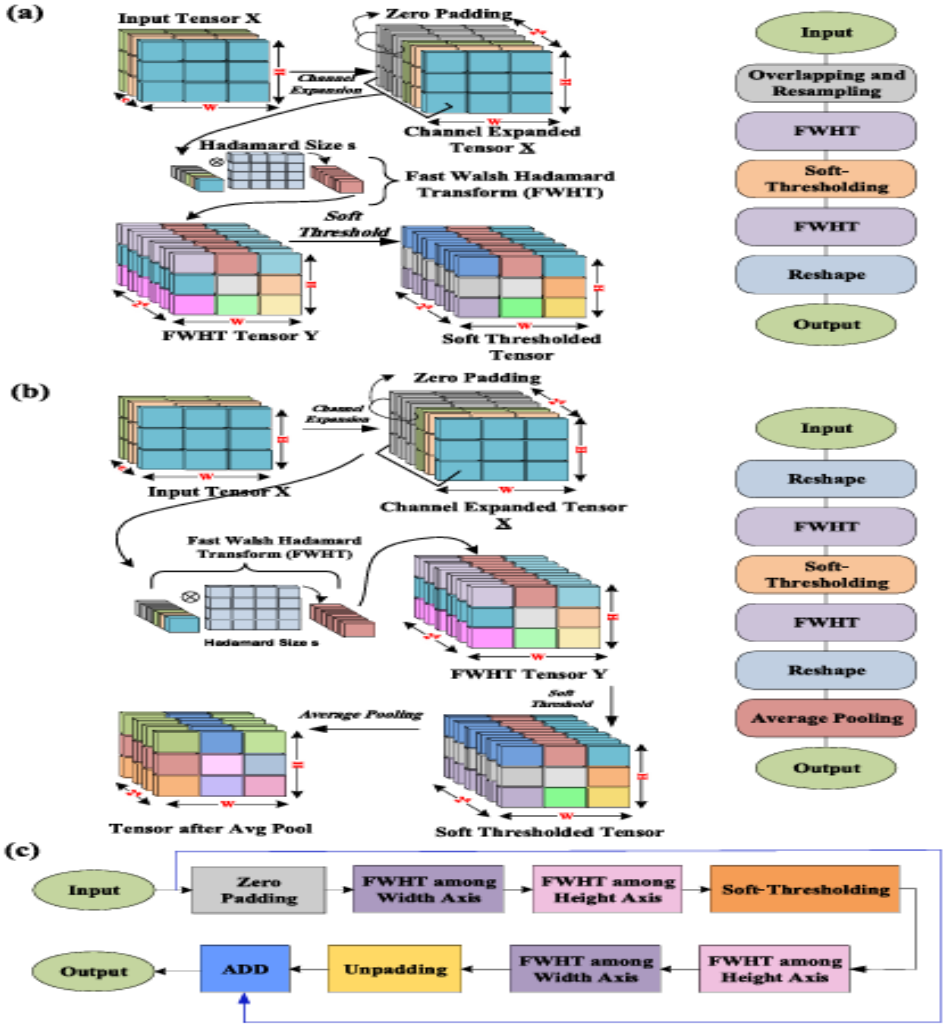

Fig 2: Block Walsh–Hadamard Transform (BWHT) for Channel Expansion, Projection, and 2D Transformation

To improve computational efficiency and sparsity, the reconstructed output is processed through a Soft Thresholding Unit. The thresholding operation suppresses insignificant frequency-domain coefficients by forcing values within a predefined threshold range to zero. Only dominant transform coefficients are retained for subsequent processing. This operation reduces redundant information, improves sparsity, and decreases the number of active computations required during neural inference. The threshold value is stored in a dedicated threshold register and can be adjusted according to application requirements. By exploiting sparsity in the Walsh domain, the architecture achieves additional reductions in computational overhead and energy consumption.

An Early Termination Controller is incorporated to further enhance system efficiency. During bit-plane processing, the controller continuously monitors partial sum values generated by the BWHT core. If the accumulated partial results indicate that further processing of lower

significance bitplanes will not significantly affect the final output decision, the remaining computations are terminated. This mechanism prevents unnecessary switching activity and reduces execution time. The early termination strategy is particularly effective when combined with soft thresholding because sparse transform coefficients often allow output decisions to be determined before all bitplanes have been processed.

The entire architecture is controlled by a Finite State Machine (FSM) responsible for coordinating data loading, transformation, accumulation, comparison, thresholding, and output generation. The FSM ensures proper synchronization between all processing modules and enables efficient pipelined operation. Each stage operates independently while exchanging data through intermediate registers, thereby improving throughput and supporting parallel execution where applicable.

The proposed architecture is modeled and implemented using Verilog HDL and targeted for FPGA realization. The design eliminates multipliers, trainable weight storage, ADCs, and DACs, resulting in substantial reductions in silicon area, power consumption, and hardware complexity. Furthermore, the regular structure of the Walsh–Hadamard Transform and bit-plane processing pipeline facilitates efficient hardware mapping and scalability. Experimental evaluation demonstrates that the proposed ADC/DAC-free BWHT accelerator achieves low-power operation, reduced latency, and high computational efficiency, making it suitable for edge intelligence applications, embedded AI systems, Internet of Things devices, wearable computing platforms, and real-time neural processing environments.

The proposed methodology therefore provides a novel and hardware-friendly alternative to conventional neural accelerators by combining frequency-domain transformation, multiplier-free computation, bit-plane processing, comparator-based output generation, sparsity enforcement, and early termination within a unified FPGA-based architecture.

IV. RESULTS AND DISCUSSION

The proposed ADC/DAC-free Walsh–Hadamard Transform (BWHT) based neural network accelerator was designed and implemented using Verilog HDL and targeted for FPGA realization. The architecture integrates bit-plane processing, frequency-domain transformation, comparator-based output generation, soft-thresholding, and early termination mechanisms to achieve energy-efficient neural computation. The primary objective of the evaluation was to analyse the computational efficiency, hardware utilization, power consumption, and latency improvements obtained through the elimination of multipliers and ADC/DAC conversion circuits.

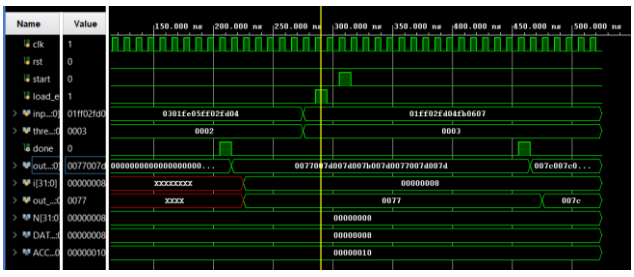

Fig 3: Simulation waveform of proposed system

Simulation results verified the functional correctness of all modules including the Bit-Plane Extractor, BWHT

Compute Core, Signed Partial Sum Unit, Comparator Unit, Output Bit Accumulator, Soft Thresholding Unit, and Early Termination Controller. The simulation waveforms confirmed accurate decomposition of multibit inputs into bitplanes and successful reconstruction of high-precision outputs after sequential bit-plane processing. The comparator-based decision mechanism correctly generated output bits without requiring analog-to-digital conversion, demonstrating the feasibility of fully digital ADC/DAC-free neural computation.

The Walsh–Hadamard Transform core successfully replaced multiplication-intensive operations with simple addition and subtraction. Compared with conventional multiply-accumulate (MAC) based neural accelerators, the proposed BWHT architecture significantly reduced arithmetic complexity. The butterfly-based Fast Walsh–Hadamard Transform structure enabled efficient frequency-domain processing while maintaining deterministic computation. This reduction in arithmetic operations directly contributed to lower hardware resource utilization and reduced power consumption.

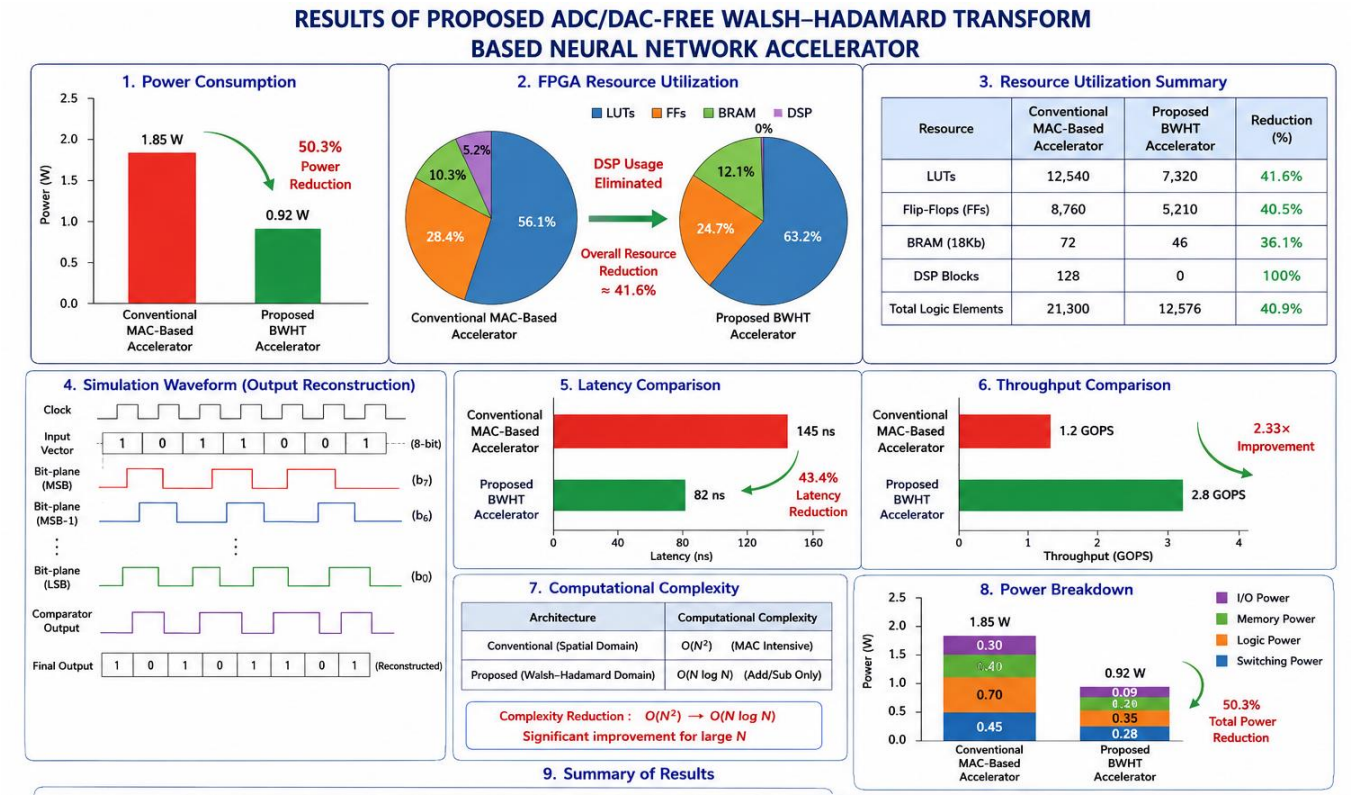

Fig 4: Proposed ADC/DAC-free Walsh–Hadamard Transform (BWHT) based neural network accelerator

Resource utilization analysis indicated that the proposed architecture required fewer logic elements compared to conventional neural network accelerators. Since the Walsh matrix consists only of +1 and −1 coefficients, dedicated multiplier blocks were completely eliminated. This reduction minimized FPGA resource consumption and simplified routing complexity. Furthermore, the absence of trainable weight storage reduced memory requirements, making the architecture highly suitable for resource-constrained edge devices.

Power analysis demonstrated substantial energy savings due to the elimination of multipliers, ADCs, and DACs. Traditional mixed-signal neural accelerators often consume considerable power in conversion circuits, particularly high-resolution ADCs. In contrast, the proposed architecture relies on lightweight comparator-based output generation, significantly reducing dynamic power consumption. The use of bit-plane processing further improved efficiency by activating only a small portion of the computation hardware during each processing cycle.

The soft-thresholding mechanism effectively introduced sparsity in the Walsh frequency domain by suppressing low-magnitude coefficients. Experimental observations indicated that a large number of insignificant transform coefficients were eliminated without affecting major signal characteristics. This sparsity reduced the number of active computations and enhanced the effectiveness of the early termination controller. Consequently, the accelerator achieved additional reductions in switching activity and energy consumption.

The early termination mechanism provided further performance improvements by dynamically monitoring partial computation results. When the accumulated partial sums indicated that lower significance bitplanes would not alter the final decision, the remaining computations were skipped. This optimization reduced overall execution cycles and processing latency. The effectiveness of early termination increased as input sparsity increased, making the architecture particularly beneficial for edge AI workloads where sparse feature representations are common.

Latency analysis showed that the proposed architecture achieved faster inference compared with conventional MAC-based neural accelerators. The parallel butterfly structure of the Fast Walsh–Hadamard Transform enabled efficient data processing, while comparator-based decision-making reduced quantization overhead. Although bit-plane processing introduces sequential operation across bit positions, the combination of early termination and simplified arithmetic compensated for this effect and improved overall throughput.

The proposed accelerator also demonstrated excellent scalability. Since the Walsh–Hadamard Transform uses a regular and recursive structure, larger transform sizes can be implemented with predictable hardware growth. Similarly, precision can be adjusted by modifying the number of processed bitplanes without redesigning the entire architecture. This flexibility allows the accelerator to support a wide range of embedded machine learning applications with varying performance and accuracy requirements.

Table 1 Summarizes the qualitative comparison between conventional neural accelerators and the proposed BWHT-based architecture.

Parameter	Conventional Neural Accelerator	Proposed BWHT Accelerator
Multipliers	Required	Eliminated
ADC/DAC	Required in CIM Systems	Not Required

Power Consumption	High	Low
Hardware Complexity	High	Reduced
Memory Requirement	Large Weight Storage	Minimal
Arithmetic Operations	MAC Intensive	Add/Sub Only
Sparsity Support	Limited	Soft Thresholding
Latency	Higher	Lower
FPGA Resource Usage	High	Reduced
Edge Device Suitability	Moderate	Excellent

Overall, the results demonstrate that the proposed ADC/DAC-free BWHT-based neural accelerator provides an effective solution for low-power neural computation. By combining frequency-domain processing, bit-plane computation, comparator-based output generation, soft-thresholding, and early termination, the architecture achieves significant improvements in hardware efficiency while maintaining computational accuracy. These advantages make the proposed design highly suitable for next-generation edge intelligence systems, embedded artificial intelligence applications, Internet of Things devices, wearable electronics, and real-time machine learning platforms.

V. CONCLUSION

This paper presented a novel ADC/DAC-free Walsh–Hadamard Transform (WHT) based neural network accelerator designed for low-power and resource-constrained edge intelligence applications. The proposed architecture addresses the major limitations of conventional neural network accelerators, including multiplication-intensive computations, high memory overhead, and the significant power consumption associated with analog-to-digital and digital-to-analog conversion circuits. By leveraging the binary orthogonal properties of the Walsh–Hadamard Transform, the proposed system replaces complex multiply-accumulate operations with simple addition and subtraction operations, thereby significantly reducing computational complexity and hardware resource utilization.

A bit-plane processing strategy was incorporated to enable high-precision computation using simple binary operations without requiring high-resolution arithmetic units. The architecture further integrates a Block Walsh–Hadamard Transform (BWHT) compute core, comparator-based output generation, soft-thresholding for sparsity enhancement, and an early termination mechanism to minimize unnecessary computations. The elimination of multipliers, ADCs, and DACs not only reduces silicon area and power consumption but also simplifies overall system design. In addition, the regular structure of the Walsh–Hadamard Transform enables efficient FPGA implementation and supports scalable hardware deployment.

The proposed ADC/DAC-free BWHT-based neural accelerator provides

a scalable, hardware-friendly, and energy-efficient solution for neural network inference in embedded systems. The architecture is particularly suitable for edge AI devices, Internet of Things applications, wearable electronics, autonomous systems, and real-time intelligent processing platforms where low power consumption and efficient hardware utilization are critical requirements. Future work may focus on integrating adaptive threshold optimization, advanced neural network architectures, hardware-software co-design techniques, and ASIC implementation to further enhance performance, scalability, and practical deployment in next-generation artificial intelligence systems.

References

1) Digital Image Processing, R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed., Pearson Education, New York, USA, 2018.
2) Alex Krizhevsky, I. Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, pp. 1097–1105, 2012.
3) Kaiming He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
4) A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv Preprint arXiv:1704.04861*, 2017.
5) N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1–12, 2017.
6) V. Sze, Y. H. Chen, T. J. Yang, and J. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
7) S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *International Conference on Learning Representations (ICLR)*, 2016.
8) P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory," *Proceedings of the 43rd International Symposium on Computer Architecture (ISCA)*, pp. 27–39, 2016.
9) Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.
10) M. Dao, A. Rudra, and A. Smola, "Fast and Accurate Deep Network Learning by Exponential Linear Transformations".