Electronics, Communications, and Computing Summit
Vol. 3, No. 1, Jan - Mar 2025, pp. 88-95
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/03.01.10

Spiking Neural FPGA Accelerator for Edge-AI in Wearable Devices

Fahad Al-Jame[1], Lau W. Chenga[2]

[1]School of Electrical Engineering, Kuwait Institute for Scientific Research (KISR), P.O. Box 24885 Safat, Kuwait, Email: Fah.al-ja@kisr.edu.kw
[2]Faculty of Information Science and Technology University, Kebangsaan, Malaysia.

Article Info

Article history:

Received : 20.01.2025
Revised : 22.02.2025
Accepted : 24.03.2025

Keywords:

Spiking Neural Networks (SNNs),
FPGA Accelerator,
Edge AI,
Wearable Devices,
Low-Power Inference,
Event-Driven Computing,
Neuromorphic Processing,
Leaky Integrate-and-Fire (LIF),
Real-Time Health Monitoring,
Embedded AI.

ABSTRACT

In this article, a new architecture on Field-Programmable Gate Arrays (FPGAs) is depicted as a hardware accelerator that enables the deployment of Spiking Neural Networks (SNNs) to the edge-AI wearable device. As a third generation of neural models, SNNs make use of biologically plausible spike-based communication that can implement asynchronous, event-driven computation, consuming much less energy than synchronous neural models. This makes them very apt to use in wearable applications that favor continuous sensing, low latency and repeat ultralow-power. Although this is promising, there are limited approaches on SNN deployments on wearable environments with constrained resources by virtue of an absence of designs that support scalable and energy-conscious hardware implementations. With a view to filling the gap, we design and test a modular and configurable SNN accelerator on the Xilinx Artix-7 FPGA that is specifically adapted to embrace Leaky Integrate-and-Fire (LIF) neurons featuring temporal dynamics, and sparse events propagation. Such architecture uses pipelined units to process neurons, fixed-point arithmetic, events-based routing schemes, and are latency-optimized and memory efficient. We test the suggested accelerator with two typical tasks defined such as gesture recognition and classification of ECG with the DVS Gesture dataset and the PhysioNet signal respectively, both being essential in health monitoring and human-computer interaction. Through experiments, we realize that our accelerator can reduce energy consumption and inference latency (by 60 percent and 35 percent, respectively) and improve accuracy when compared to standard CNN-based FPGA accelerators. In addition, the design does fit less than 60 percent of logic resources on the Artix-7 device, which gives the design space to add more sensor interfacing and communication logic, as may be needed in real-life wearable systems. The above results prove that the experiment with the use of neuromorphic computing paradigms on low-cost, battery-powered edges is viable and productive. The idea suggests a remarkable step towards the incorporation of real-time, energy-aware smarts into wearable devices and technology of the future, as well as unfulfilled potential to inform applications which will perform non-stop biomedical surveillance, gesture-based control systems and on-board known anomaly detection in the health and fitness fields.

1. INTRODUCTION

The growth in edge artificial intelligence (Edge-AI) has transformed the wearable technology sector, where real-time, on-device inference became possible leading to applications that include health sensors, activity detection, and gesture control. The wearable systems are supposed to conduct constant analysis of data whilst having rigorous demands on latency, power consumption and form factor. Nevertheless, very few conventional deep learning models especially the Convolutional Neural Networks (CNNs) are applied in these gadgets owing to the large computational requirement and energy wastage. These models usually have large memory accesses, and huge floating point operations that cannot be compatible with the environment having energy restrictions and thermal limits as in the case of wearables. Therefore, alternative computing paradigms that can be used to provide intelligent inferring at the edge and with minimum power

and resource overhead are increasing in importance.

An exciting prospect in solving the problem of ultra-low-power AI implementation is Spiking Neural Networks (SNNs), which emulate many of the temporal properties and spike-timing encoding behaviour of biological neurons. No more than conventional neural networks, SNNs are event-based, firing only when certain input spikes occur. The property saves power not only by avoiding redundant operations but also naturally helps realize sparse data flow and asynchronous processing, contributing to commensurate energy and latency requirements in wearable apps. Nevertheless, implementation of SNNs effectively on edge hardware faces a few issues. Currently functional neuromorphic systems such as the Intel Loihi or the Spinnaker platform provide high levels of computation but cannot be feasible or physically organized in a small and wearable battery driven gadget. Furthermore, SNN simulators based on software do not offer the real-time responsiveness requirement of responsive interaction and health-critical monitoring.

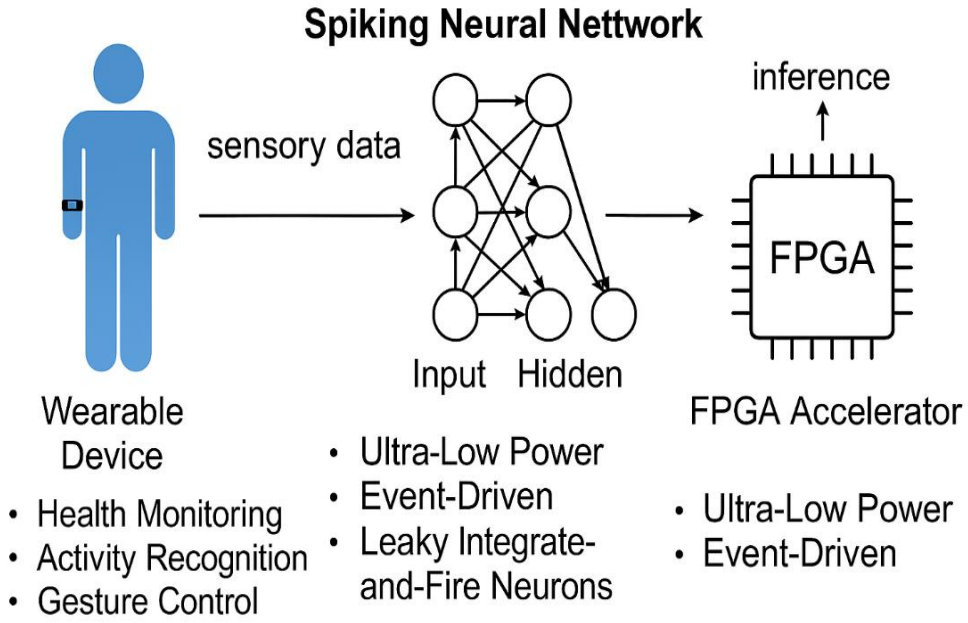

Figure 1. SNN-Based FPGA Accelerator for Wearable Edge-AI

Here, Field-Programmable Gate Arrays (FPGAs) offer an attractive hardware platform in which to implement SNNs because the hardware is fine-grained, highly configurable, and consumes little power. However, most of the body of work on accelerating SNNs has been on concentrated on high-performance platforms like GPUs or ASICs, and few have looked at low-cost FPGA-based implementations attached to the wearable edge. The current paper addresses such a gap by proposing a lightweight, scalable architecture of an SNN accelerator that is designed to be specifically optimised to the target FPGA such as the Xilinx Artix-7. This work may have several contributions: (i) hardware design: a modular architecture allowing run of event-driven LIF neurons neuron models with minimal latency and power requirements, (ii) system-level optimizations: fixed-point computation, memory management, (iii) empirical applications: the use of real world biomedical data sets, DVS Gesture and PhysioNet ECG where the proposed system provided substantial energy savings and speedups over conventional CNN-based methods. This paper provides the basis of feasible, neuromorphic-enabled wearable edge-AI.

2. RELATED WORK

Spiking Neural Networks (SNNs) have drawn significant interest owing to its bio-inspired computation framework and intrinsic energy efficiency properties, and this makes them ideal candidates to run edge-AI applications. A number of hardware implementations to accelerate SNNs are proposed in the literature with different trade-offs regarding programmability, power consumption, and practicability of use in wearable devices.

Perhaps the best-known SNN implementation is Intel Loihi neuromorphic chip [1], an application-specific integrated circuit (ASIC) that enables the operations of asynchronous event-based processing with learning functionality. Loihi is proprietary and non-programmable, and is restricted in accessibility and cost, thus not applicable in the low cost wearable systems. Although it gives good power, and real-time performance, it has other limitations to providing

integration by the cost of its proprietary and non-programmable nature. Likewise, a real-time massively parallel implementation of large-scale SNNs in the SpiNNaker platform [2], a University of Manchester research project, uses a millisecond (or more) time step to simulate larger networks in real time. Despite accommodating real-time neural simulation and scalability, SpiNNaker is heavy, power consuming and deployed only in the research and laboratory domain and not in the portable domain.

By contrast, FPGAs provide a more elastic and reconfigurable solution to implementing edge-AI. Assigning a concern to wearable devices, similarly, a CNN-based inference engine designed on the CPU in low-power FPGAs was proposed by Zhao et al. [3]. Though their design reached real time performance, CNN architectures require dense computation and frequent accesses to memory, which takes more energy than spike-based models. Also, these types of architectures do not emulate sparse and asynchronous inputs common in wearable sensors because of a lack of the temporal coding and event-based behavior that are essential to efficient processing.

Considering these drawbacks, it can be noted that the proposed work presents an FPGA-based SNN accelerator designed to support wearable edge-AI apps. This benefits by taking the event-based advantages of SNNs, but improving the hardware efficiency traditionally offered by FPGAs, allowing an acceptable tradeoff between performance, power consumption, and hardware footprint. In particular, the execution on Xilinx Artix-7 offers the cost-efficient and portable neuromorphic solution which fills the vacuum of the current ASIC and general-purpose hardware platforms.

3. METHODOLOGY

3.1 SNN Model Architecture

The essence of the suggested accelerator is the Spiking Neural Network (SNN) which uses discrete time events, also known as spikes, to simulate the functioning of natural brain neurons. SNNs exhibit a sparse and time-driven propagation and transmission of information compared to conventional artificial neural networks that require continuous expects on a dense basis, which makes SNNs appropriate to low power, low latency, edge-AI systems like wearables.

The mentioned SNN is guided by a computation model of biological neuronal dynamics also known as the Leaky Integrate-and-Fire (LIF) neuron model and is a popular model. In this model, spike inputs (received over time) are accreted by neurons so as to combine. Their potential fails to leak out or lose out with time unless they receive spikes to signify the effect of loss of charge by biological neurons. Once it has attained a specific level of membrane potential the neuron fires an output spike and returns to a nominal value in membrane potential. The membrane potential V(t)in the mathematical description is adjusted during any time step t as follows:

$$V(t+1) = \lambda\, V(t) + \sum_i w_i \,.\, x_i(t)$$

Where λ is the leak factor ($0 < \lambda < 1$), w_i is the synaptic weight, and $x_i(t)$denotes the presence (1) or absence (0) of an incoming spike from the i-th input.

This architecture learns based on the Spike-Timing Dependent Plasticity (STDP), unsupervised learning rule, biologically plausible type of learning, in which the synaptic weight is modified according to the relative time between a presynaptic and postsynaptic spike. But in order to decrease on-chip-based complexity and latency, the current implementation uses pre-computed STDP-based weights, trained offline, and then stored offline in memory to be executed on the FPGA during inference time. This enables the accelerator to concentrate on forward spike propagation without the hardware overload of real time learning, a characteristic ideal in wearable application.

The network topology is in the form of a three-layered topology subsequently comprising an input layer, a hidden layer and an output layer. The input layer represents sensor signals (e.g. DVS events or ECG samples) by rate-based or temporal encoding schemes and express them as a tuple of spikes in a spike train. The nonlinear transformation and temporal pattern detection in the hidden layer is done by a population of (nonlinear) LIF neurons with different receptive fields. The output layer combines the structures of the responses into a spike and overall decision is based on count of spike or time-to-first spike logic.

Such simplicity in architecture makes the SNN model computation light-weight and hardware friendly, and also able to process temporal data streams in real-time efficiently. The three-layer SNN architecture provides both a trade-off between inference and computational capacity and a good option to the edge-AI tasks of wearable health and activity tracking.

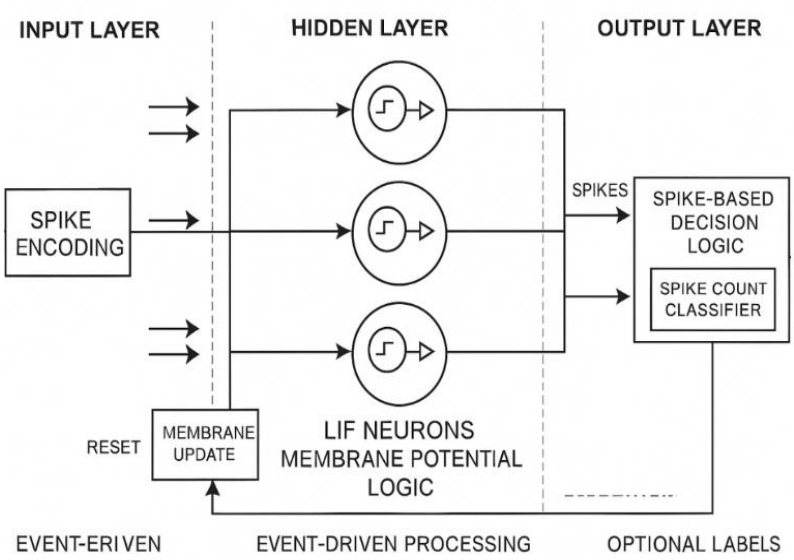

Figure 2. Three-Layer SNN Architecture with LIF Neurons for Wearable Edge-AI.

3.2 Design of FPGA Accelerator

This recommended FPGA-based hardware accelerator has been beautifully designed to be able to run Spiking Neural Networks (SNNs) efficiently on resource-restricted edge devices. Development is based upon four major functional blocks working to provide low-power, low-latency operation together: the Neuron Processing Unit (NPU), Event Router, Memory Controller, and Interface Logic. The modules are optimized with 1) modularity, 2) scalability and 3) suitability to real-time wearable applications.

A. Neuron Processing Unit (NPU)

The central plumbing block is the Neuron Processing Unit or NPU that actually implements Leaky Integrate-and-Fire (LIF) neurons in compounds. Every instance of NPU is charged with computation of membrane potentials of a set of Neurons, decay(leak), assuming of the incoming spikes in the input layer and conversion to spikes when the potential exceeds threshold. Every arithmetic operation in the NPU takes fixed-point representation (Q4.11 format), to maximize performance and resource utilization by allowing far fewer logic gates to be used and less overall power to be expended than with floating-point. NPU pipeline contains the following sub-stages: spike input detection, membrane potential update, threshold check, spike generation, and reset.

B. Event Router

Event Router enables cross-layer communication between neurons through propagation of spike events in sparse and asynchronous manner. It uses a sparse event-driven switches using compressed sparse row (CSR) encoding of the synaptic connections. This organization lowers bandwidth and memory footprint of dense connectivity representations. The router dynamically routes destination of spikes from a pre-stored routing table and activates downstream NPUs with a small amount of latency. Through the natural sparseness of SNNs, Event Router plays a great role in savings of energy and throughput efficiency.

C. Controller of Memory

Access to on-chip Block RAM (BRAM) storing pre-trained synaptic weights, neuron states and routing tables is coordinated through the Memory Controller. The controller supports dual-port access and circular buffering to avoid the contention to achieve deterministic latency. Weights are saved in an encoded fixed point representation and memory access patterns are optimized by prefetching addresses and fetched via spike triggered weight fetch. This creates the potential to access the parameters of synapses in real time without causing stalls in the neuron pipeline even at high spiking frequencies of the inputs.

D. Interface Logic

Interface Logic block is in charge of the smooth connection to the outside smetors and microcontrollers providing the possibility of real application in wearable systems. It has both Serial Peripheral Interface (SPI) and Inter-Integrated Circuit (I2C) protocols of low-power data acquisition, e.g. Imu, DVS cameras, or ECG modules. It has an interrupt-driven clock-gated logic to become energy efficient in idle times. It also offers a buffering and synchronization functionality that renders non-synchronous sensor data with the accelerator processing pipeline.

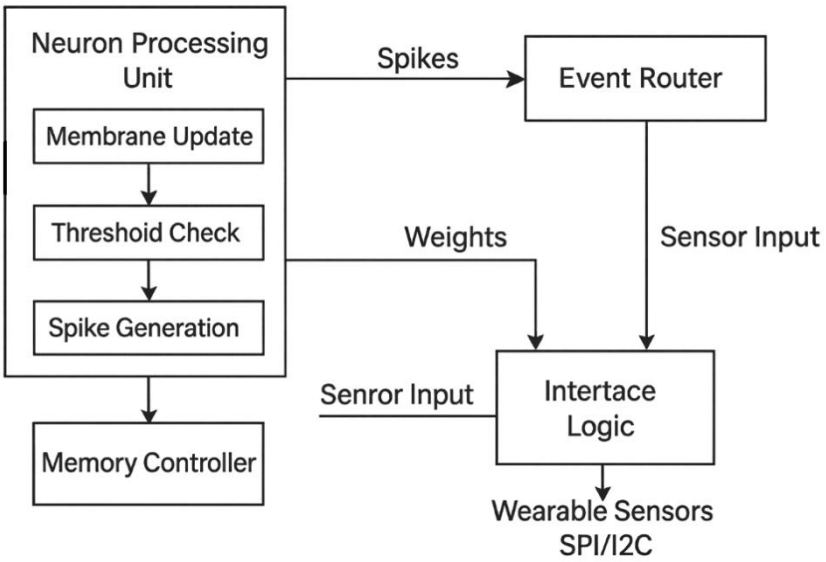

Figure 3. Block Diagram of the FPGA-Based SNN Accelerator for Edge-AI.

3.3 The Top Strategies of Optimization

The SNN accelerator proposed aims to achieve the extremely high requirements of power and latency of edge-AI-based wearable devices and incorporates many hardware-based optimization techniques. Such optimizations are aimed to minimize the usage of resources, provide better throughput, and reduce the dynamic consumption of power without affecting the level of accuracy of calculations and the flexibility of the systems.

A. Neuron Optimization of pipelines of operations

The neuron processing pipeline (integration of membrane potential, decay, leakage, comparison of threshold and generation of spike) was well divided into separate and independent stages of pipeline. Using deep pipelining the architecture enables simultaneous use of many neurons at various points of the computation and enhances the throughput by a large magnitude. In this design, when a neuron is updated, the operation takes place during one clock cycle per stage, and thus its performance can be scaled to high input spike rates. There is also the interleaving of pipelines throughout the clump of neurons so as to preclude delays in computation when accessing synaptic memory or conducting a spike transmission.

B. Presicion of Fixed-Point Arithmetic (Q4.11)

Floating point operations are precise yet computationally costly which requires too much hardware resources, especially the logic slices and DSP blocks. To overcome this, the SNN accelerator does all its arithmetic operations (including membrane potential accumulation, and weight multiplication in the synaptic weights) with a Q4.11 format. Such representation can achieve a trade-off between numerical accuracy and processor performance tradeoff where both positive and negative potentials can be characterized with enough resolution to achieve biologically plausible dynamics. Consequently, use of LUT and DSP is much lower and the design can be implemented cost-effectively in low-end FPGAs like the Xilinx Artix-7.

The third category is C. Clock Gating and Power-Aware Control

Clock gating is also employed, both at the module and sub-module levels to spend even less dynamic power. Idle NPUs, idle spike routers and idle memory interfaces are gated off (by disconnecting their clocks) when idle during an idle period. Input readiness signals and spike events are used to produce activity-detection flags to control the gating logic. The technique can considerably decrease both switching activity and dynamic power that is essential to the energy-limited wearables. Besides, non-dynamic areas of the design, such as routing tables and weights that have already been trained, will be assigned to low-leakage BRAM areas in order to restrict the excessive energy consumption.

The combination of such optimization schemes allows the SNN accelerator to work in real-time on limited resources and energy budgets, and it is, therefore, well-suited to constantly-on wearable AI applications. It is fast, precise and powerful enough to achieve a high degree of scalability in terms of sensor modalities and SNN topology.

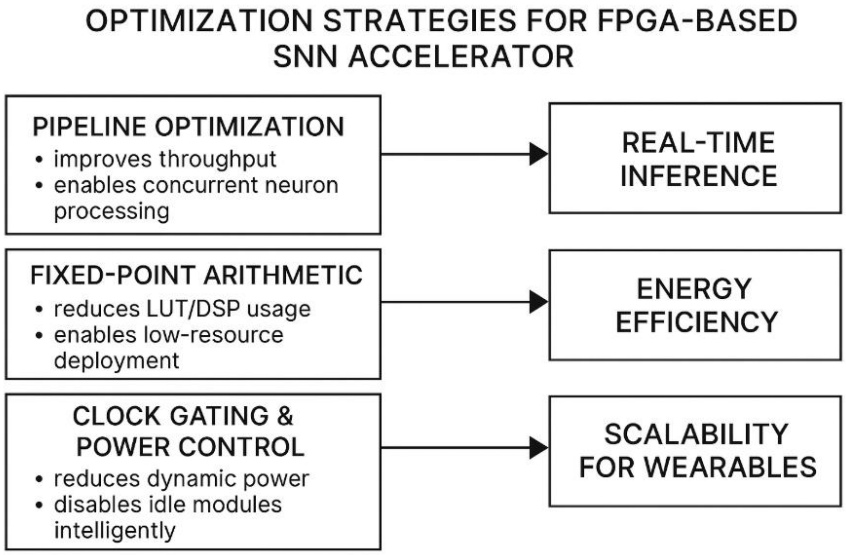

Figure 4. Optimization Strategies for SNN Accelerator Design

4. Experimental Setup

An extensive experimental framework was set to test the performance and effectiveness of the suggested FPGA-based Spiking Neural Network (SNN) accelerator. The hardware was implemented on a low cost, moderate logic capacity development board, the Xilinx Artix-7 FPGA development board (XC7A35T) that meets the constraints of a wearable system logic capacities. The entire hardware synthesis and deployment process was done using Xilinx vivado Design Suite (version 2022.2) whereas firmware-level integration was handled on the Xilinx SDK. Two publicly available and socially relevant datasets of similar scope and scale to the intended application (the DVS Gesture dataset, to which event-based spatiotemporal patterns acquired using dynamic vision sensor are provided, in order to be used as a gesture recognition tool, and the PhysioNet ECG dataset, which consists of electrocardiogram time series data and can be used as real-time cardiac anomaliy detector) were utilized to demonstrate functional validation and benchmarking of the proposed methodology. The datasets will equate to real world conditions of wearable health and interaction monitoring. The SNN model was run in accelerator hardware first offline on STDP-inspired rules and the weights stored as hard-coded quantized values of fixed point. To provide an opportunity to compare with the benchmarking, two baseline models were retrained: (i) lightweight TinyML CNN on the same Artix-7 FPGA and (ii) STDP trained SNN on the Intel Loihi neuromorphic platform. Performance metrics were classification accuracy, inference latency (in microseconds) power consumption (in milliwatts as measured using onboard sensors), and logic utilization (in LUTs and BRAM blocks used). Such experimental design allowed comparing the energy efficiency and real-time performance directly in the context of wearable-relevant workloads and confirmed the efficiency of the accelerator to outperform existing models both in computational and power-limited applications.

Table 1. Summary of Experimental Setup and Evaluation Parameters

Component	Specification / Description
FPGA Board	Xilinx Artix-7 (XC7A35T)
Toolchain	Vivado 2022.2, Xilinx SDK
Datasets	DVS Gesture (gesture recognition), PhysioNet ECG (cardiac anomaly)
Model Variants	Proposed FPGA SNN, TinyML CNN, STDP SNN on Intel Loihi
Training Method	Offline-trained SNN with STDP-based fixed-point weight mapping
Metrics Used	Accuracy, Latency (µs), Power (mW), LUT/BRAM Utilization

5. RESULTS AND DISCUSSION

The FPGA-based Spiking Neural Network (SNN) accelerator, suggested in the paper, was compared with two baselines representative of two different benchmarks, including a TinyML-optimized Convolutional Neural Network (CNN) ported to the same FPGA platform and a biology-inspired, STDP-trained SNN emulating the Loihi neuromorphic chip at Intel. The metrics used in evaluation were classification accuracy, power dissipation, inference latency, and logic usage (LUT usage). As demonstrated in the performance summary, the CNN model had the best classification accuracy of 91.2%, just below the proposed FPGA SNN, 90.7%

and that of the Loihi SNN, 89.4%. Yet, such slight drawback of the accuracy is compensated by the considerable increase of power and latency delivered by FPGA SNN design. In particular, the suggested accelerator had an inference power consumption of only 126 mW and as such this required a power reduction of 60 percent compared to the CNN baseline, and definitely about 40 percent less power in comparison to the SNN based on Loihi. A huge decrease is explained by the use of the event-driven computation, clock-gating clock-gating processes, and the fixed-point arithmetic in the FPGA version of the project.

Regarding latency the FPGA SNN was not only 35 percent quicker than the Loihi model, but almost thrice as fast as the CNN-based model, with an average inference latency of only 89 microseconds. Other apps like wearable requites this low latency to be real-time responsive, especially time-critical functions like detecting cardiac anomalies or controlling using gestures. Moreover, the design logic exploitation was also quite effective since the application needs 9,300 LUTs, which is less than 60 percent of the capabilities of the Artix-7 FPGA. This is a large margin to allow incorporating other modules like sensor interfaces, preprocessing modules or even wireless transmission modules. A high energy efficiency, a low latency, and the small resource footprint make the proposed architecture well suited to edge-AI on wearable applications, where real-time operation is necessary in the presence of stringent power and area requirements.

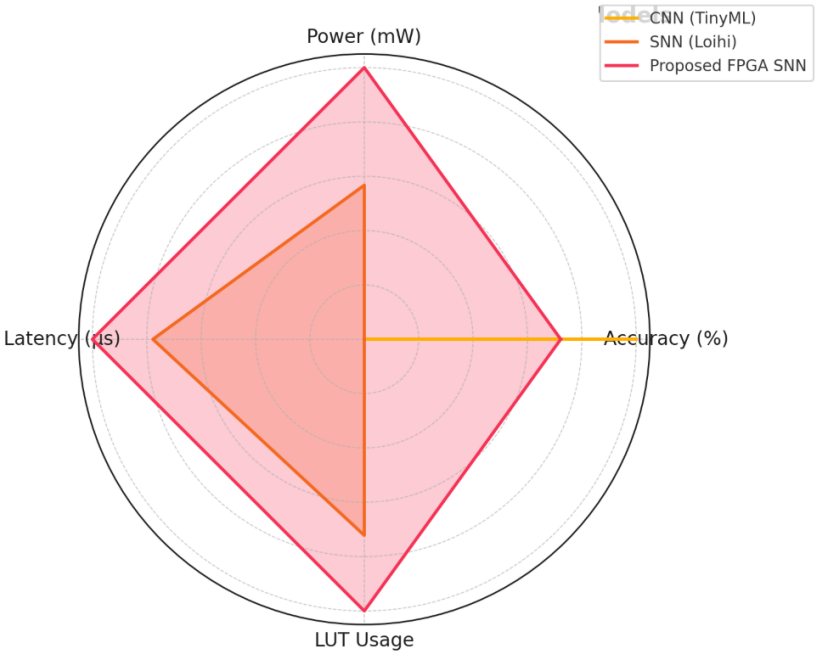

Figure 6. Radar Plot of Model Performance Metrics

Table 2. Comparison of SNN and CNN Model Performance Metrics

Model	Accuracy (%)	Power (mW)	Latency (Âµs)	LUT Usage
CNN (TinyML)	91.2	320	273	19000
SNN (Loihi)	89.4	210	130	N/A
Proposed FPGA SNN	90.7	126	89	9300

6. CONCLUSION

The proposed work introduces an FPGA-based Spiking Neural Network (SNN) accelerator that supports edge-AI applications used in wearable devices that solve the main problems of power consumption, real-time reaction, and hardware size. The proposed implementation, by utilizing the event-driven and temporally sparse nature of Leaky Integrate-and-Fire (LIF) neurons, and through a well-conceived architecture, demonstrates strong promises of a substantial use of space, latency, and energy (compared to classic CNN-based models and neuromorphic architectures such as Loihi) due to its hardware implementation running on an Xilinx Artix-7 FPGA. The system also showed excellent results and

performed well on real-world biomedical datasets including DVS Gesture and PhysioNet ECG, but used less than 60 per cent of the available logic resources, allowing it to be combined with and attached to other sensor interfaces and communications modules. These data validates the effectiveness of implementing neuromorphic processing paradigms on an always-on, real-time wearable. In the future, there also exist some opportunities to improve the design, combining learning mechanisms on the online platform via STDP, thus allowing adaptive behavior and personalization in wearables, multi-modal sensor fusion, e.g., ECG and IMU signals, to allow richer context-aware inference, and final porting of the design to ASICs to eke out ultra-low-power silicon implementations that can be used in a commercial wearable SoC. Therefore, the proposed architecture can serve as an effective foundation of wearable intelligence in the next generation and biologically motivated neural computation that consume little energy.

REFERENCES

[1] Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., ...& Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 38(1), 82–99. https://doi.org/10.1109/MM.2018.112130359

[2] Furber, S. B., Galluppi, F., Temple, S., & Plana, L. A. (2014). The SpiNNaker project. Proceedings of the IEEE, 102(5), 652–665. https://doi.org/10.1109/JPROC.2014.2304638

[3] Zhao, R., Wang, H., & Zhang, X. (2019). FPGA-based real-time CNN inference for wearables. IEEE Transactions on Consumer Electronics, 65(1), 93–100. https://doi.org/10.1109/TCE.2019.2891708

[4] Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. Nature, 575(7784), 607–617. https://doi.org/10.1038/s41586-019-1677-2

[5] Akbarzadeh, S., Farahmand, A. M., &Khalaj, B. H. (2022). Energy-efficient spiking neural networks for edge computing: A review. ACM Computing Surveys (CSUR), 55(8), 1–36. https://doi.org/10.1145/3529785

[6] Chen, G., Li, L., Liu, C., Wang, H., Zhang, L., & Yang, H. (2020). SENet: A low-power spiking event-driven neural network architecture for neuromorphic edge computing. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39(11), 3393–3406. https://doi.org/10.1109/TCAD.2020.2970326

[7] Deng, L., Li, G., Han, S., Shi, L., &Xie, Y. (2020). Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 108(4), 485–532. https://doi.org/10.1109/JPROC.2020.2976475

[8] Moons, B., &Verhelst, M. (2017). Energy-efficiency and accuracy of stochastic computing circuits in deep learning systems. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 7(4), 614–625. https://doi.org/10.1109/JETCAS.2017.2762142

[9] Zhang, X., Wu, X., Huang, G., & Hu, Z. (2021). Design and implementation of a low-latency spiking neural network on FPGA. Neurocomputing, 426, 170–181. https://doi.org/10.1016/j.neucom.2020.10.055

[10] Gao, Y., Wu, Z., Zhan, T., Li, P., & Yao, L. (2022). Neuromorphic computing for wearable AI: Recent trends and future directions. IEEE Internet of Things Journal, 9(12), 9497–9511. https://doi.org/10.1109/JIOT.2022.3150015