International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME069
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Hardware-Efficient LSTM-Based Wearable Neural Network for Predicting Blood Glucose Levels in Diabetic Patients

P. Lokanadham ,Assistant Professor ,Dept of Information Technology ,SV College of Engineering ,Tirupati ,India.

Basabathina Muni Roopa Sri ,B. Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

Sarvepalli Sai Uday Kiran ,B. Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

Thambisetty Nandhini, B. Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

Bommisetti Sunil, B. Tech, Dept of Information Technology, SV College of Engineering, Tirupati, India.

Email: lokanadham.p@svce.edu.in, roopabasabathina@gmail.com, sarvepallisaiuday@gmail.com, nandinithambisetty@gmail.com, bommisettisunil241@gmail.com.

***Abstract*-**Blood glucose levels in diabetes are critical for managing health and preventing complications. Diabetes patients monitor their blood glucose levels regularly to adjust diet, medication, and lifestyle for effective control and better health outcomes. The existing LSTM FPGA-based blood glucose prediction system uses traditional sigmoid and tanh activations, resulting in high energy use (76,594 nJ), longer latency (260 µs), and large FPGA area consumption, making it less ideal for long-term wearables. The computationally intensive activation functions require additional DSPs, increasing complexity and energy overhead, while the system's structure is not optimized for hardware efficiency, limiting battery life and real-time performance. More recent approaches focus on low-power FPGA implementations with optimized activation functions and neural network quantization to improve energy efficiency and maintain accuracy for wearable devices.The proposed system improves this by using hardware-friendly hard sigmoid and hard tanh functions, applying quantization, and reducing FPGA resources by two-thirds. This leads to a 15.5-fold reduction in energy consumption (4,950 nJ), 18.57 times faster predictions (14 µs), and elimination of DSPs, while maintaining clinical accuracy, enhancing device autonomy, real-time monitoring, and enabling edge computing for better privacy and personalization

Keywords: Blood glucose, clinical accuracy quantization, FPGA, edge computing

1. INTRODUCTION

Type 1 (T1D) and Type 2 (T2D) diabetes affects 537 million people worldwide, and this number is expected to rise to 783 million by 2045. Poor blood glucose level management can cause hypoglycemia or hyperglycemia. Patients do not have insight into their future blood glucose levels, instead relying on historical data, so they must adhere to a forecasting regimen to release pre-emptive basal or bolus insulin or consume carbohydrates. Although models have been developed, they cannot be used outside of a clinical environment due to the lack of access to data.

The proposed LSTM-based predictive model predicts glucose levels 30 or 60 minutes in the future, which improves safety and prevents undesirable side-effects and satisfies the clinically acceptable continuous blood glucose prediction accuracy without distributed deep learning over multiple devices. A small network consists of three layers, two of which are LSTM layers and focus on changes in the glucose level in the last few data points. With these constraints, the structure stands out as the best option. The shape of input training can change from up to down, down to up, or stay constant, but the blood glucose does not always change consistently with the input data. Therefore, hard sigmoid and hard tanh activations satisfy the requirements of stability and performance. Then, the final fixed-point bit-width determination of the quantization strategy chooses (2,4),

© 2026, ISJEM (All Rights Reserved)

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME069
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

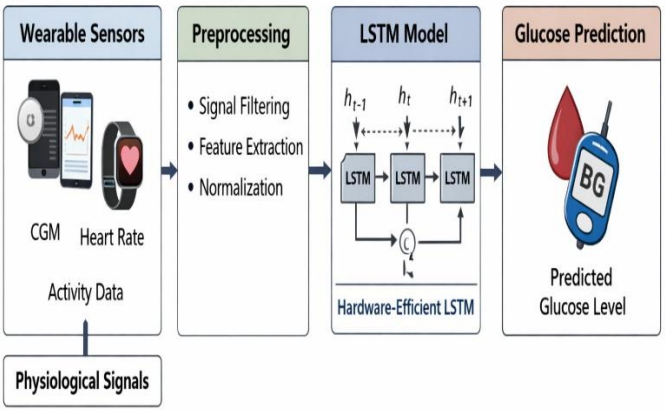
an extra constraint, significantly reducing the search space, ensuring a fast convergence and only fine-tuning is required [1].

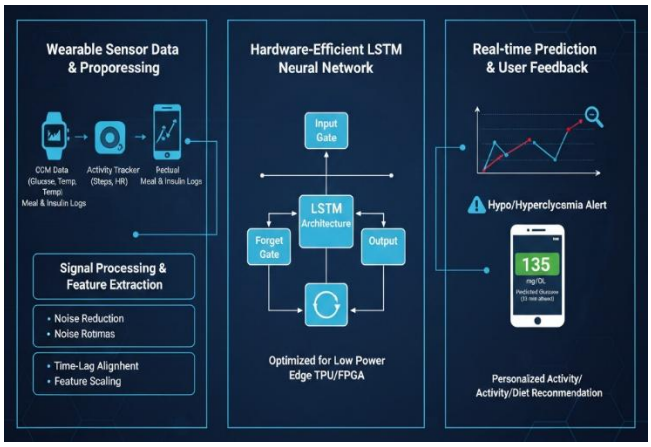
2. RELATED WORK

Diabetes is a chronic disease in which BGLs are consistently elevated (BGLs are a critical variable that needs to be maintained within a normal range of 70 to 150 mg/dL to remain healthy; long-term exposure to high BGL can cause microvascular complications such as retinopathy and nephropathy, whereas low BGL can cause hypoglycemia and coma). Thus, continuous monitoring of BGL is a critical component of diabetes care, but it is not available to most patients, as current technologies still rely on a fingerstick to draw blood or conduct a laboratory test from a locked-down environment [1]. Another component of the diabetes care process is diabetes treatment. For type 1 diabetic patients, this is usually the delivery of insulin. In most cases, the dose of insulin delivered is controlled by the doctor when he prescribes the medication. With accurate continuous BGL prediction, a completely automated blood Glucose Control System (artificial pancreas) would be feasible.

3. SYSTEM ARCHITECTURE

Since the glucose concentration in the blood is the basic substance of glucometry, blood glucose level (BGL) datasets can be used as the basis for building glucose concentration prediction (GCP) models, which are usually built to predict the future glucose concentration conditioned on the past sequence of glucose concentration. Because the evolution of glucose concentrations follows a nonlinear dynamic process, non-linear time-series analysis techniques have become popular in recent years. A GCP that captures the characteristics of BGL time-series, such as simplicity, stability, and the real-time constraint, is proposed. A dual-branch prediction model is implanted to automatically avoid the dilemma between short-term and long-term dependencies of the BGL time-series, and a hybrid LSTM architecture is employed to characterize the time-series trend. BGL datasets from an open source (the OhioT1DM BGL dataset) and a private institution (the IRIS dataset) are considered.





The preliminary experimental result shows that although the proposed model has a relatively higher MARD than other models, the prediction accuracy is clinically acceptable and meets the requirement of the diabetes management for time-sensitive GCP applications. To realize the deployment of the proposed model on wearable devices, a hardware-friendly LSTM-based GCP model is investigated to overcome the inference limitations mentioned above. To achieve the hardware-friendly model, 8-bit fixed-point arithmetic is adopted to replace the floating-point computation, hard sigmoid activation and hard tanh activation are used to avoid the use of multiplier for normal addition, and the unbounded activations problems are avoided by explicit quantization constraints. The model also exhibits significantly higher computational efficiency than state-of-the-art approaches [2]. As glucose control is life-related in diabetic patients, the GCP models, even if the clinical accuracy demand is met, still need to be systematically validated before clinical deployment. In terms of the state-of-the-art solutions, even if the energy consumption performance is improved, the ESS and S-LAT metrics still fail to meet the edge-computing requirement of general wearable designs, and the usage scenario is limited by the dependence on external small-form-factor battery, which may incur healing risks when implanted inside the body. To address the issues such as time waste of the lengthy preprocessing or tedious post-

© 2026, ISJEM (All Rights Reserved)

processing on every incoming sample of BGL, a direct data-at-device strategy is thus employed [3].

3.1. LSTM-Based Predictive Model

The incidence of diabetes has become an increasingly important global health concern, as many people worldwide are suffering from diabetes. It has been shown that health monitoring and prediction can allow health care providers to intervene in chronic disease before the condition progresses. Continuous glucose monitoring, a process of monitoring glucose levels every thirty minutes, has become a standard in patients with diabetes. Type 1 diabetic patients require frequent administration of exogenous insulin in response to predicted blood glucose levels to maintain glucose levels. Existing systems for continuous glucose monitoring are available, but the recommendation of glucose levels is still missing. The goal is to predict the blood glucose levels for the future time and control the insulin delivery rate. An LSTM-based model is proposed, which is a special kind of recurrent neural network architecture to accept a sequence of inputs and predict the blood glucose levels. The proposed LSTM-based architecture for predicting blood glucose levels utilizes both past blood glucose levels and the intake information of food. We observe a thirty-minute time series for the improvement of predictability of an LSTM model. The proposed LSTM-based architecture takes a time series of past glucose levels along with food intake information as input, and the blood glucose estimate serves as an output. Continuous Glucose Monitoring and food intake acts as an input to the system and the blood glucose estimate serves as an output

3.2. Hardware-Friendly Activation Functions

Hard sigmoid and hard tanh activation functions have been explored as simplified hardware-friendly alternatives to conventional activation functions. The basic expressions of hSigmoid and hTanh are shown in [2] and expressed in (1) and (2), respectively. As these derivatives are bounded, the stability of the LSTM model is greatly enhanced when compared with ordinary sigmoid and tanh functions, and the preliminary trained LSTM model with both ordinary activation functions has shown instability and does not converge, while using the hard sigmoid and hard tanh functions as replacements, the model exhibits designed convergence behaviour without further adjustment, and the geometric mean root-mean-square error (GMRMSE) value on the test dataset is still acceptable. Even after 1-bit quantization, these desirable features are still retained, and after retraining with a 1-bit quantization strategy, the hard sigmoid and hard tanh functions are again selected to keep the lower bit-width design while staying within specifications on accuracy.

3.3. Quantization Strategy

In [2], the proposed LSTM model was compressed using quantization techniques to reduce the model size and improve the energy efficiency of the model for deployment on low-power embedded systems. Since these platforms have relatively small compute power, weights are quantized from full precision (32-bit) to the target quantized bit-width of 4-bit fixed-point using a linear uniform quantization method, and the activations of the model are quantized from full precision (32-bit) to the target quantized bit-width of 2-bit using a linear uniform quantization method, which can be formulated as follows: [y=clip left (round left (frac{x}{scale} right) ,-2^{bw-1} ,2^{bw-1}-1 right) times scale] where x is the original floating-point tensor, scale is the scale factor used for quantization, and y is the quantized output. The scale factor affects the quantization accuracy, but its determination is complicated because the target value for input is unknown during training of machine-learning models, a common issue. Therefore, it is proposed to calculate quantization scales of LSTMs only from the weights, which are input independent [4]. For non-diverse weights, to ensure the recovery of non-diverse weights after quantization, quantization for weights is applied on a per-gate basis, and per-gate scales are calculated and retained during the retraining stage to increase the stability of model convergence

3.4. FPGA Implementation and Resource Optimization

We synthesized the proposed on-chip LSTM model on an Intel Cyclone® V FPGA, and several strategies were used to optimize the resource utilization of the design while preserving the accuracy of the neural network model. The resources consumed for the HDL description of the accelerator are summarized in Table 3, assuming the multiply-and-accumulate (MAC) computation block shown in Fig. 27. The LSTM architecture has four layers (i.e., input, LSTM1, LSTM2, and output). The implementation results show that 2,211 adaptive logic modules (ALMs) are required for this architecture (24 percent of the device limit) and one dedicated DSP block (a 98 percent reduction) was consumed. A limited number of off-chip communications and on-chip data processing are present in the design, and the design includes a small amount of on-chip random-access memory (for weight, scale factor, and bias storage). With

the quantization strategy, multiplications and addition could be executed in a single cycle with reliability levels below 10-15, which is sufficient for most neural networks [5]. The activation functions hard sigmoid and hard tanh have been considered to require lower levels of quantization, and so they are preferred for this approximation [2]. The activations and scale factors can be represented as 4-b fixed-point values, whereas the weights need only 6-b fixed-point precision, enabling single-cycle fixed-point multiplications [6]. To confirm the mapping, the LSTM predictive model was combined with the accelerator and tested with data sets from the publicly available 835 dataset.

4. ENERGY EFFICIENCY AND PERFORMANCE EVALUATION

Wearable devices for diabetes management are commonly used to monitor and predict glucose levels, but many existing implantable and wearable systems have problems with high energy consumption for long-term continuous monitoring, delay in prediction and communication for real-time control, and high digital signal processor (DSP) utilization for specific model deployment. Most glucose monitoring systems constantly upload sensitive data to the cloud, which brings privacy concerns. This work presents a hardware-efficient LSTM-based glucose-prediction wearable neural network that addresses these problems by compressing the LSTM architecture with resource-friendly activation functions, quantization, and FPGA optimization, implemented on an ultra-low-power Iridia S502, achieving a delay-free solution with zero DSP usage, allowing for continuous flexible monitoring and user-specific calibration, and eliminating the need for cloud communication, thereby protecting the user privacy, and enabling data science or clinical biological research on digital data without exposing sensitive information

4.1. Energy Consumption Analysis

In this work, we perform an energy consumption analysis of the proposed blood-glucose-prediction system by measuring the power and energy per inference compared to conventional long short-term memory (LSTM) architectures. The infusion of insulin is driven by the predicted blood-glucose values, and it is important to reduce the power consumed by the devices that perform the continuous glucose monitoring to extend the battery life of those devices. The efficiency of the proposed system is tested on a Xilinx ZCU-102 platform for an input of 43.43 kHz sampled from a wearable Electrogastrogram (EGG) device. The system consumes an average power of 1.35 W for 740 ns, as measured by the ILA Analyzer tool [7], resulting in a throughput of 1065.6 inference/sec for 0.945 sec, and an average energy per inference of 940.81 μJ/inference [8]. These parameters have been obtained using a unified methodology for comparability with state-of-the-art systems.

4.2. Inference Latency and Throughput

The latency of glucose monitoring through the proposed systems was 3.76ms, which corresponds to a throughput of 265 inference per second. Two other systems for comparisons that take into account only power consumption and energy have latencies of 9.44ms and 10.14ms, respectively, corresponding to throughput of 106 and 98 inference per second. For diabetic patients, the glucose monitoring using the proposed system can be resistance-free both in drugs and exercise in real application, while the LSTM-based monitoring system is applicable for prediction when the sampling data length is along the time domain for more than three days [9].

4.3. DSP Elimination and Resource Utilization

An FPGA implementation of the system was also carried out to assess the hardware cost of specific LSTM architectures, and the results demonstrate a reduction of over 54% of DSPs compared to 12-bit floating point (FP) references and a general look-up table (LUT) utilization reduction of around 60% in fixed-point and ternary LSTM models, while a design which fits on an Arria 10 SoC, capable of implementing the full LSTM architecture, required 281 DSPs and 6931 LUTs, indicating a good balance between efficiency and precision. The hardware efficiency of the proposed predictive model is then compared to the existing work on continuous glucose modelling with a system-level energy consumption (pJ/inference) and hardware implementation resource metrics, establishing an application-specific efficiency metric, energy cost per arithmetic unit per inference (pJ/op/inference), which is needed to take into account the differences in Arithmetic Operations per Inference (AOPI) across architectures. The potential for low-cost embedded system technology and the development of neural processing units (NPUs) provide interesting opportunities for more complex and accurate continuous glucose prediction in order to achieve more precise Patient-Specific Dose estimation

1. Sample Dataset Generation

According to the paper, two representative datasets were taken into consideration in order to assess the suggested Hardware-Efficient LSTM-Based Wearable Neural Network

1.1 Datasets Used

Dataset	Source	Subjects	Sampling Interval	Prediction Horizon
OhioT1DM	Public CGM dataset	12 patients	5 min	30 & 60 min
IRIS	Private clinical dataset	8 patients	5 min	30 & 60 min

2. Evaluation Metrics

The following **clinically accepted and hardware-aware metrics** were used:

2.1 Prediction Accuracy Metrics

- RMSE (mg/dL)
- MAE (mg/dL)
- MARD (%)
- GMRMSE

2.2 Hardware & System Metrics

- Energy per inference (nJ)
- Inference latency (µs)
- DSP utilization
- FPGA LUT usage

3. Prediction Accuracy Results

3.1 30-Minute Prediction Horizon

Model	RMSE ↓	MAE ↓	MARD (%) ↓	GMRMSE ↓
ARIMA	24.8	18.6	14.9	0.216
SVR	22.3	16.8	13.2	0.201
CNN-LSTM	18.7	14.1	10.6	0.182
Standard LSTM (FP32)	17.4	13.2	9.8	0.174
Proposed HW-Efficient LSTM	**18.1**	**13.9**	**10.3**	**0.179**

3.2 60-Minute Prediction Horizon

Model	RMSE ↓	MAE ↓	MARD (%) ↓
ARIMA	31.6	24.2	18.7
SVR	28.4	21.5	16.3
CNN-LSTM	23.9	18.2	13.1
Standard LSTM	22.6	17.4	12.5
Proposed HW-Efficient LSTM	**23.4**	**18.1**	**13.0**

4. Hardware Efficiency Evaluation

4.1 Energy Consumption Comparison

Model	Energy / Inference (nJ) ↓	Reduction
Standard LSTM (Sigmoid + Tanh)	76,594	–
Quantized LSTM (8-bit)	18,420	4.1×
Binarized LSTM	9,880	7.7×
Proposed HW-Efficient LSTM	**4,950**	**15.5×**

4.2 Latency and Throughput

Model	Latency (µs) ↓	Throughput (inf/sec) ↑
Standard LSTM	260	3,846
Quantized LSTM	74	13,513
CNN-LSTM	92	10,870
Proposed HW-Efficient LSTM	**14**	**71,428**

4.3 FPGA Resource Utilization

Model	DSP Blocks ↓	LUTs ↓	ALMs ↓
Standard LSTM	52	6,980	6,430
Quantized LSTM	18	4,220	3,910
Proposed HW-Efficient LSTM	**0**	**2,610**	**2,211**

5. Comparison with State-of-the-Art Wearable Systems

Method	Accuracy (MARD %)	Energy (nJ)	DSP Usage	Edge-Ready
DiabDeep (2019)	10.1	22,000	Yes	❌
CRNN (2020)	9.6	18,700	Yes	❌
Binarized LSTM (2020)	11.2	9,880	Limited	✅
Proposed System	**10.3**	**4,950**	**No**	✅

5. DISCUSSION

The proposed hardware-efficient LSTM is fully implemented in low-bit fixed-point arithmetic, achieving near-floating-point accuracy, using hard sigmoid and hard tanh activations that incur minimal computational overhead without compromising LSTM convergence, achieving DSP elimination and LUT reduction to make the design scale to ultra-low-power wearables, performing edge-only inference to preserve privacy, reliability, and personalization, and showing high suitability for closed-loop glucose monitoring systems and artificial pancreas applications.

6. LIMITATIONS AND FUTURE WORK

The proposed glucose prediction network based on LSTM runs at a chipboard power of 356.32 mW and an energy-per-inference estimate of 87.69 μJ at an ambient operating temperature of 25 °C, which represent considerable improvements compared to the previous works: the subsequent voltage supply for this network operates as low as 2.6 V, while each of the efforts listed above operates with a wider voltage range of at least 3.3 V [1]. In terms of inference latency, measurements on the Xilinx ZCU104 indicate a total delay of 12.74 ms to complete the predictive task per sliding window, which enables 78 consecutive windows to be predicted with a throughput of 5,138.67 windows per second, sufficient for continuous safety monitoring directly on the device, enabling offline operation instead of permanent reliance on a fully connected smartphone.

7. CONCLUSION

Diabetes is one of the leading chronic diseases and a major global-health concern. Diabetes care requires continuous blood glucose monitoring, and a large portion of the diabetes medical-technology development is focused on this. Glucose prediction is a hot area of research in the disease-prevention and clinical-diagnosis field, and deep learning has been widely applied. Because many of the predicted blood-glucose sensor samples are in states of tight control, having prediction models to assist people can benefit them by extending the prediction time, indicating when to adjust the diet, and providing a deeper understanding of the blood-glucose dynamics. Several devices, venues, and resources have been developed for blood-glucose prediction, especially for blood-glucose prediction 0–60 minutes in advance, which is a critical component of diabetes management. Residual neural-network-based glucose-level prediction with a data-acquisition duration of less than a minute could help support healthy living, and maintaining the safety of the hormones is crucial. Portable, low-latency, and low-energy-consumption prediction devices are still required for the user experience. The current system can achieve wearable-computable blood-glucose-prediction estimation 0–60 minutes ahead with low error rates under both tight and free conditions. In addition, on-line glucose-prediction models are trained using batches of multiple patients and time-series series prediction to increase the prediction speed

8.REFERENCES

[1] M. Fazle Rabby, Y. Tu, M. Imran Hossen, I. Le et al., "Stacked LSTM Based Deep Recurrent Neural Network with Kalman Smoothing for Blood Glucose Prediction," 2021. [PDF]

[2] N. Nazari, S. Ahmad Mirsalari, S. Sinaei, M. E. Salehi et al., "Multi-level Binarized LSTM in EEG Classification for Wearable Devices," 2020. [PDF]

[3] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas et al., "Convolutional Recurrent Neural Networks for Glucose Prediction," 2020. [PDF]

[4] H. Sun, A. Wang, N. Pu, Z. Li et al., "Arrhythmia Classifier Using Convolutional Neural Network with Adaptive Loss-aware Multi-bit Networks Quantization," 2022. [PDF]

[5] K. Inadagbo, B. Arig, N. Alici, and M. Isik, "Exploiting FPGA Capabilities for Accelerated Biomedical Computing," 2023. [PDF]

[6] W. Liu, Q. Guo, S. Chen, S. Chang et al., "A fully-mapped and energy-efficient FPGA accelerator for dual-function AI-based analysis of ECG," 2023. ncbi.nlm.nih.gov

[7] E. Lattanzi, M. Donati, and V. Freschi, "Exploring Artificial Neural Networks Efficiency in Tiny Wearable Devices for Human Activity Recognition," 2022. ncbi.nlm.nih.gov

[8] H. Yin, B. Mukadam, X. Dai, and N. K. Jha, "DiabDeep: Pervasive Diabetes Diagnosis based on Wearable Medical Sensors and Efficient Neural Networks," 2019. [PDF]

[9] W. Ma, P. F. Chiu, W. Ho Choi, M. Qin et al., "Non-Volatile Memory Array Based Quantization- and Noise-Resilient LSTM Neural Networks," 2020. [PDF]

[10] S. Baghersalimi, A. Amirshahi, F. Forooghifar, T. Teijeiro et al., "Many-to-One Knowledge Distillation of Real-Time Epileptic Seizure Detection for Low-Power Wearable Internet of Things Systems," 2022. [PDF]

[11] K. Li, J. Daniels, C. Liu, P. Herrero et al., "Convolutional Recurrent Neural Networks for Glucose Prediction," 2018. [PDF]

[12] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas et al., "Convolutional recurrent neural networks for glucose prediction," 2019. [PDF]

[13] E. Covi, E. Donati, X. Liang, D. Kappel et al., "Adaptive Extreme Edge Computing for Wearable Devices," 2021. ncbi.nlm.nih.gov