IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 10, 2019, accepted November 24, 2019, date of publication November 27, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956179

A Neural Network-Based ECG Classification Processor With Exploitation of Heartbeat Similarity

JIAQUAN WU[1], FEITENG LI[1], ZHIJIAN CHEN[1], YU PU[2], AND MENGYUAN ZHAN[3]
[1]Institute of VLSI Design, Zhejiang University, Hangzhou 310027, China
[2]Alibaba DAMO Academy, Sunnyvale, CA 94085, USA
[3]The Affiliated Hospital of Qingdao University, Qingdao 266000, China

Corresponding author: Zhijian Chen (chenzj@vlsi.zju.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61801425.

ABSTRACT This paper presents a neural network based processor with improved computation efficiency, which aims at multiclass heartbeat recognition in wearable devices. A lightweight classification algorithm that integrates both bi-directional long short-term memory (BLSTM) and convolutional neural networks (CNN) is proposed to deliver high accuracy with minimal network scale. To reduce energy consumption of the classification algorithm, the similarity between consecutive heartbeats is exploited to achieve a high degree of computation reuse in hardware architecture. In addition, neural network compression techniques are adopted in the procedure of inference to save hardware resources. Synthesized in the SMIC 40LL CMOS process, the prototype design has a total area of 1.40 mm^2 with 186.2 kB of static random-access memory (SRAM) capacity. Based on the simulation, this processor achieves an average energy efficiency of 3.52 GOPS/mW under 1.1 V supply at 100 MHz frequency. Compared with the design without computation reuse, the proposed processor provides a speedup by 2.58x and an energy dissipation reduction by 61.27% per classification. This work is a valuable exploration of neural network based design for long-term arrhythmia monitoring in daily life.

INDEX TERMS Computation reuse, electrocardiogram (ECG) processor, input similarity, neural networks.

I. INTRODUCTION

Cardiovascular arrhythmia is a common disease that may occur suddenly and become life-threating if not treated properly [1]. Wearable arrhythmia monitoring devices that are based on electrocardiogram can record and analyze long-term physiological signals in real time, which greatly improve the life quality and survival rate of patients.

The high demand of wearable arrhythmia monitoring devices leads to various electrocardiogram (ECG) processors, which target at improving energy efficiency and achieving high detection accuracy. Bayasi *et al.* [2] present a low-power ECG processor for predicting ventricular arrhythmia by adopting a naive Bayes classifier. However, the prediction accuracy is limited to 86% for ventricular arrhythmia detection only. Ieong *et al.* [3] presents a real-time compression processor for extracting valuable ECG information. In spite of the high compression ratio of this design, the intrinsic latency and high energy consumption of data transmission are problems yet to be solved. Chen *et al.* [4] process ECG signals with a weak-strong hybrid classifier, which acquires a nice balance between classification accuracy and energy dissipation, but it is ineffective for multiclass heartbeat classification. Xu *et al.* [5] realizes energy-efficient multi-class heartbeat classification with the granular resampling method and adaptive speculative mechanism. However, the classification accuracy remains unsatisfactory, especially for the ventricular class (only 78.67%).

Overall, these ECG processors that use the traditional classification methods have two aspects to be improved: *i)* Most previous works only distinguish ventricular heartbeats (V) from normal heartbeats (N), and few of them accomplish high accuracy heartbeat recognition for all the standard Advancement of Medical Instrumentation (AAMI) [6] classes; *ii)* The fixed manual features employed by the traditional classification methods are not sufficient when treating different

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

patients with various types of arrhythmia [7], as the connotative characteristics underlying the original ECG signals may be ignored.

Neural network algorithms [8]–[11] achieve high accuracy in multiclass ECG recognition and demonstrate good generalization capability for different patients. As the ECG signals vary significantly among different patients, deep neural network topologies are required to gain the high classification accuracy, but the large memory usage and computation complexity present significant challenges to the implemention of the neural network based ECG processors in wearable devices.

To reduce the heavy workload load of these neural network based algorithms, the patient-specific means [7], [12], [13] are proposed to extract the best possible features of individuals by exploiting relatively simple network architectures. For each patient, a personal neural network model is trained by the relatively small common training data from all patients and the specific data collected from that patient. Once the individual network is well trained, the following heartbeat classification and ECG monitoring are based on that model for specific patient. The patient-specific method dramatically decreases the demand for computation and storage hence becoming more applicable to wearable devices.

In this paper, we propose a neural network based processor with improved energy efficiency for patient-specific ECG classification. The heartbeats are classified into five classes (N, supraventricular ectopic beat (S), V, fusion of a ventricular and a normal beat (F), or unknown beat type (Q)) strictly following the AAMI guidance [6]. The main contributions are as follows:

i) A lightweight neural network based ECG classification algorithm with high recognition accuracy is proposed by combining both the bi-directional long short-term memory (BLSTM) [14] and convolutional neural networks (CNN), which better satisfies the demand of limited energy consumption and hardware resources on wearable devices.

ii) The high degree of similarity between successive heartbeats is utilized to achieve computation reuse on hardware architecture, which greatly speeds up the network inference and improves the energy efficiency.

iii) Network compression techniques, including weight quantization and parameter precision reduction, are adopted to reduce the storage of parameters and shrink the scale of processor with negligible loss on classification accuracy.

The rest of this paper is organized as follows: The proposed ECG classification algorithm is described in Section II. Section III presents the hardware architecture design of the processor. Section IV evaluates the processor implementation. Section V concludes this paper.

II. ALGORITHM DESIGN

The ECG classification algorithm plays an important role in optimizing the processor. This section presents our low footprint neural network with integrated BLSTM and CNN.

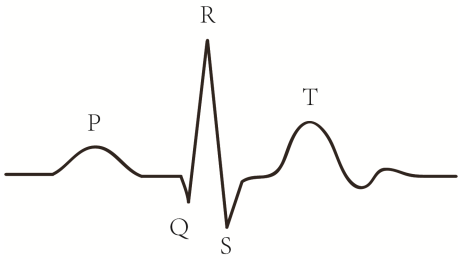


FIGURE 1. One typical heartbeat in ECG signal with significant details, including the P, Q, R, S, and T waves.

A. BACKGROUND

One typical heartbeat in ECG signal is shown in Fig. 1. The significant clinical details include the P, Q, R, S, and T waves that represent the electrical activities of cardiac muscle [15]. The P wave describes the Atria depolarization (Atrial Contraction), the T wave shows the depolarization of ventricles (Ventricular relaxation), and the QRS complex represents ventricles depolarization (Ventricular contraction) [16]. The study of ECG morphology is helpful for the optimization of heartbeat classification algorithms.

B. INTEGRATING BLSTM AND CNN

To reduce the cost of ECG processors, neural network based classification methods are required to minimize their model scales while extracting more effective features. The network topology with integrated BLSTM and CNN is proposed to fulfill the purpose. Since the CNN extracts the morphological features of ECG signals significantly [12] and the BLSTM is superior in handling sequence tasks [17], the combination of them can implement efficient feature extraction, and thus achieving a high classification accuracy with smaller amount of parameters and less computation load.

The proposed ECG classification algorithm is described in Fig. 2. The original ECG waveforms of two signal channels are first normalized for removing external noises, such as baselines wander [18], and then they are segmented into single heartbeats according to the average RR interval D of each patient. The heartbeat for the BLSTM usage comprises the sample points of 1/3 of the average RR interval before the R peak and 2/3 of the average RR interval after the R peak. As the process of ventricular repolarization (causing T waves) is longer than that of sinoatrial node depolarization (causing P waves) [15], the imbalanced periods before and after the R peak are adopted for the completion of P and T waves. The heartbeat for the CNN usage comprises the sample points of 4/5 of the average RR interval both before and after the R peak. The extra sample points are applied to enrich the details of P waves and T waves, which helps to distinguish N and S. This presented adaptive segmentation method maintains the uniformity of heartbeats from different patients, which helps to extract effective features. To balance the workload and classification performance of our network, the segmented heartbeats are resampled to 128 points for the BLSTM and

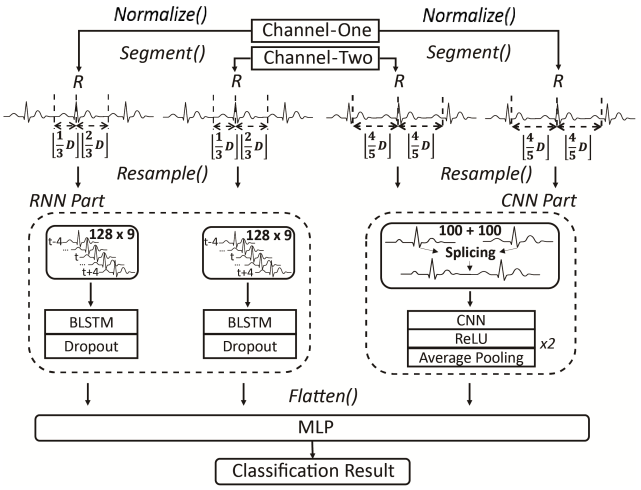


FIGURE 2. The architecture of the proposed ECG classification algorithm. The ECG waveforms of two signal channels after baseline wander removing are first segmented into single heartbeats and then resampled as the inputs of BLSTM and CNN. The locations of R waves are adopted as reference to extract heartbeats with adaptive lengths according to the average RR interval *D*.

100 points for the CNN using the linear interpolation method. The q-th point $V^A(q)$ after re-sampling can be calculated with the p-th point $V^B(p)$ before re-sampling by eq. (1) and (2),

$$V^A(q) = (V^B(p+1) - V^B(p)) \times (q \times \frac{P-1}{Q-1} - p) + V^B(p) \quad (1)$$

$$p \le q \times \frac{P-1}{Q-1} \le p+1 \quad (2)$$

The proposed neural network based classifier is composed of the parallel recurrent neural network (RNN) part and CNN part. The RNN part has two sub-models that are made up by the cascaded BLSTM layer and dropout layer [19], and each sub-model deals with the single channel ECG inputs. The previous 4 heartbeats, the current heartbeat, and the following 4 heartbeats make up the time steps for one inference to extract the beat-to-beat correlation from context with the superiority of BLSTM, which is helpful for utilizing the important long-term features, such as heart rate variance. Employing two smaller sub-models in parallel instead of a larger one helps to decrease the network scale of the RNN part. The CNN part consists of two cascaded convolutional (CONV) layers that are followed by the rectified linear unit (ReLU) and average-pooling layer. The segmented heartbeats of two channels are concatenated as the CNN inputs, which helps to extract the detailed morphological features. The parallel models extract connotative features from the original ECG signals, and the outputs of both RNN part and CNN part are linked as the inputs of the multi-layer perception (MLP) to produce the final prediction results. The parameters of neural network layers are listed in Table 1, which is a detailed description of the lightweight algorithm.

C. ALGORITHM PERFORMANCE

The performance of the proposed ECG classification algorithm is tested on the popular MIT-BIH arrhythmia database [20], and the selection of common dataset DS1 and patient-specific dataset DS2 is the same as other works [7], [12], [13]. The DS1 includes record 100, 101, 103, 105, 106, 108, 109, 111, 112, 113, 114, 115, 116, 117, 118, 119, 121, 122, 123, 124, and the DS2 includes record 200, 201, 202, 203, 205, 207, 208, 209, 210, 212, 213, 214, 215, 219, 220, 221, 222, 223, 228, 230, 231, 232, 233, 234. There are 75 N, 75 S, 75 V beats and all 13 F, 7 Q beats from the DS1 making up the representatives common training data. The first five minutes of each record in DS2 work as patient-specific training data, and the remaining minutes of each record in DS2 are as testing data. The confusion matrix is listed in Table 2, which shows the high classification accuracy for N, S, V, and F beats.

TABLE 1. Parameters of neural network layers.

CNN Parameter				
Layer Name	Output Size	Kernel/Pool Size	Padding	Stride
CNN Input	[200x1]	-	-	-
CONV L0	[100x32]	16	SAME	2
POOL L0	[25x32]	4	SAME	4
CONV L1	[13x32]	16	SAME	2
POOL L1	[4x32]	4	SAME	4
BLSTM Parameter				
Layer Name	Output Size	Dropout Keep Prob		
BLSTM Input	[9x128]	-		
LSTM FW L0	[9x128]	0.1		
LSTM BW L0	[9x128]	0.1		
MLP Parameter				
Layer Name	Input Size	Output Size		
MLP	[9*128*4+4*32]	[5]		
Computation				
Total Operation		9.34 MOPs		
CNN Operation		0.29 MOPs		
BLSTM Operation		9 MOPs		
MLP Operation		0.05 MOPs		

TABLE 2. Confusion matrix of the proposed algorithm.

		Algorithm Label					
		n	s	v	f	q	Sen
Reference	N	40799	555	242	137	10	97.7
	S	329	1879	117	3	3	80.6
	V	295	37	4408	58	3	91.8
	F	57	6	88	459	1	75.1
	Q	5	1	1	0	1	12.5

As the V and S beats are more attentive in the clinic according to the AAMI guidance [6], their results are presented in more details. Four standard metrics, accuracy (Acc), sensitivity (Sen), specificity (Spe), and positive predictivity (Ppr) are applied to evaluate the classification performance. Two widely used statistical measures of F1 score and G score adopted in [7] are listed to synthetically estimate the Sen and Ppr. An overall performance comparison between the proposed algorithm and other advanced patient-specific methods [7], [12], [13], [21] are shown in Table 3. Our classification performance of V beats is comparable to existing methods, while the classification performance of S beats is more outstanding than others. The Sen, F1 and G of the S beats

TABLE 3. Performance comparison with other advanced algorithms.

Method	VEB						SVEB					
	Acc	Sen	Spe	Ppr	F1	G	Acc	Sen	Spe	Ppr	F1	G
Kiranyaz [13]	98.6	**95.0**	98.1	89.5	92.2	92.2	96.4	64.6	98.6	62.1	63.3	63.3
Zhai [12]	98.6	93.8	99.2	92.4	93.1	93.1	97.5	76.8	98.7	74.0	75.4	75.4
Saadatnejad [7]	**99.2**	93.0	**99.8**	**98.2**	**95.5**	**95.5**	98.3	66.9	**99.8**	**95.7**	78.0	80.0
Amirshahi [21]	97.9	80.2	99.8	97.3	88.0	88.3	-	-	-	-	-	-
Proposed	98.6	91.8	99.4	94.0	92.9	92.9	**98.3**	**80.6**	99.1	82.4	**81.5**	**81.5**

TABLE 4. Performance comparison of different network architectures.

Network Architecture	VEB						SVEB					
	Acc	Sen	Spe	Ppr	F1	G	Acc	Sen	Spe	Ppr	F1	G
Proposed Model	98.6	91.8	99.4	94.0	92.9	92.9	98.3	80.6	99.1	82.4	81.5	81.5
Model with independent BLSTM	98.1	88.4	99.1	91.7	90.0	90.0	97.3	56.6	99.3	79.4	66.1	67.0
Model with independent CNN	98.2	91.4	99.0	90.4	90.9	90.9	97.6	77.4	98.6	72.7	75.0	75.0
Model with independent enlarged CNN	97.6	92.0	98.2	84.9	88.3	88.4	97.1	80.2	97.9	65.3	72.0	72.4
Model with unified BLSTM	98.4	90.8	99.2	92.2	91.5	91.5	97.8	75.7	98.8	76.4	76.0	76.0

TABLE 5. Performance comparison of different BLSTM input sequence lengths.

Sequence Length	VEB						SVEB					
	Acc	Sen	Spe	Ppr	F1	G	Acc	Sen	Spe	Ppr	F1	G
3	98.5	92.2	99.2	92.2	92.2	92.2	97.4	70.5	98.8	73.9	72.1	72.2
5	98.3	92.5	98.9	90.4	91.4	91.4	98.0	74.9	99.1	80.6	77.6	77.7
7	98.6	91.9	99.3	93.4	92.6	92.6	98.1	77.5	99.1	81.2	79.3	79.3
9	98.6	91.8	99.4	94.0	92.9	92.9	98.3	80.6	99.1	82.4	81.5	81.5
11	98.5	91.8	99.2	92.9	92.3	92.4	98.1	75.7	99.2	83.3	79.3	79.4

classification are the best among all. It is worth mentioning that the [21] proposes an efficient low-power design for neural network based ECG classification. However, it only achieves four-class (N, V, F, and Q) heartbeat recognition, and the classification accuracy of V can still be improved further.

The proposed algorithm only demands approximately 9.34 MOPs for each detection owing to the novel network topology design with effective feature extraction, which is relatively lightweight when compared with other works (for example, the networks in [12] demands approximately 100 MOPs). The lower requirement for the amount of computations is beneficial for the hardware implementation.

D. NETWORK ARCHITECTURE DISCUSS

1) COMBINING BLSTM AND CNN

The topologies of BLSTM and CNN in our classification model are separately employed to validate the proposed network architecture, and the scale of MLP layer is adjusted accordingly. As Table 4 describes, the classification performance deteriorates significantly when adopting either the independent BLSTM or CNN. As both detailed morphological features and sufficient sequential features are needed to distinguish SVEB, the big accuracy loss is introduced in the SVEB classification (15.3% loss of F1 for independent BLSTM and 6.4% loss of F1 for independent CNN). Since the recognition of VEB depends more on the morphological features, the accuracy loss of independent CNN is smaller than that of BLSTM.

For further verifying the proper network scale of independent CNN, we enlarge the filter number of each layer from 32 to 128 and adjust the convolutional layer number from 2 to 3. However, as Table 4 describes, the Sen of both SVEB and VEB increase while the Ppr of them decrease more. Considering the performance of whole network completely, the scale of CNN are finally set as Table 1.

2) BLSTM TOPOLOGY OF PARALLEL SUB-MODELS

The proposed classification network architecture with two parallel BLSTM sub-models is compared with that using one unified BLSTM model. The original two-channel BLSTM input heartbeat sequences are concatenated to form the inputs of the unified model. The network scale of one unified BLSTM model is doubled than that of two parallel sub-models. The classification performance of two different BLSTM architecture designs is described in Table 4. Since the normal QRS complexes are usually prominent in the channel-one signals and the ectopic beats are more typical in the channel-two signals [20], the design of two sub-models can significantly improve the classification performance on both VEB and SVEB. The experimental results prove the effectiveness of the proposed parallel sub-model architecture that better utilizes the ECG signals of both channels.

3) BLSTM INPUT SEQUENCE LENGTH

The classification models with various input heartbeat sequence lengths (3, 7, 9, and 11) of the BLSTM are evaluated on the testing dataset. In general, as describes in Table 5, the F1 and G of both VEB and SVB classification roughly improve with the increase of the input sequence length (under 9), while the performance deteriorates when the input

sequence length is adjusted from 9 to 11. Since a longer sequence does not improve the classification performance, the proposed length of 9 (previous 4 heartbeats, the current heartbeat, and following 4 heartbeats) is the optimal choice.

III. HARDWARE ARCHITECTURE DESIGN

The hardware architecture that takes full advantage of computation reuse and network compression helps to remove the redundant computations of our ECG classification algorithm, and thus greatly reducing the energy consumption with negligible accuracy loss.

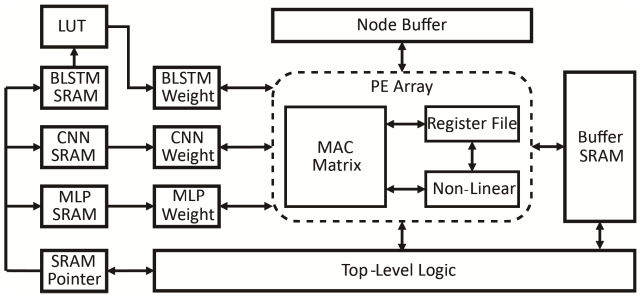


FIGURE 3. The system overview of the proposed processor. The weights of BLSTM, CNN and MLP layers are stored in SRAMs and read out to corresponding weight buffers independently. The PE array accomplishes the MAC operations and non-linear operations with the reuse of the MAC Matrix. The intermediate results are stored in the Buffer SRAM. The top-level logic is responsible for the control of the whole system.

The system overview of the proposed processor is shown in Fig. 3. The weights of BLSTM, CNN and MLP layers are all stored in SRAMs and read out to corresponding weight buffers independently during the network inference. The processing element (PE) array is reused among the computation of different topologies, which implements 16 multiply and accumulate (MAC) operations across the network layer nodes and weights per clock cycle. The non-linear unit in PE array processes the sigmoid and hyperbolic tangent functions in BLSTM using the piecewise linear function. The intermediate results generated by the PE array are stored in the buffer SRAM. The top-level logic handles the scheduling of different function units.

A. COMPUTATION REUSE DESIGN FOR INPUT SIMILARITY

As the ECG waveforms are nearly periodic signals in most situations due to the regular pattern of cardiac activities, the continuous heartbeats show significant similarity between each other. Since the BLSTM inputs are composed of the successive heartbeats to be multiplied with the same weight matrix in our design, a large amount of intermediate computation results are the same and can be reused between time steps. The processor can be significantly optimized by the computation reuse based architecture, as the calculation of BLSTM occupies the majority portion in the proposed algorithm (according to Table 1).

The process of input similarity based computation reuse in BLSTM is described in Fig. 4. For the current inference,

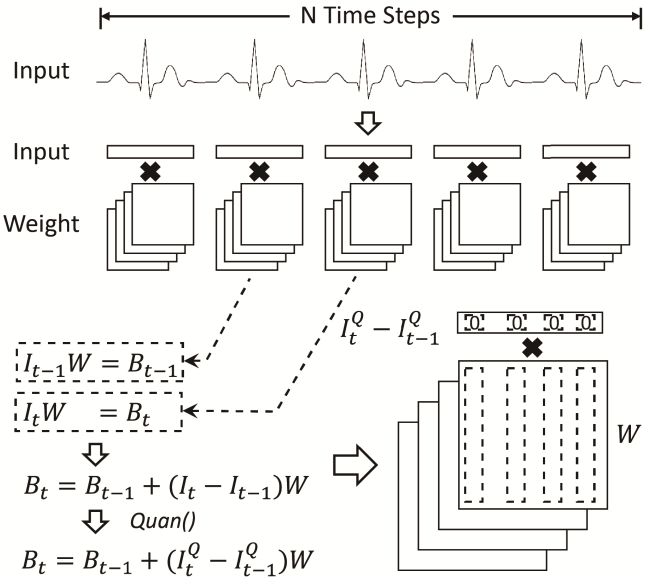


FIGURE 4. The input similarity is utilized to achieve the reuse of matrix-vector MAC results. The MAC results of the current time step B_t can be calculated by the MAC results of the last time step B_{t-1} and the matrix-vector products $(I_t - I_{t-1})W$. As there are a lot of zero values in $(I_t - I_{t-1})$ after quantization owing to the input similarity, the matrix-vector calculations can be significantly simplified in hardware design by zero-skipping.

the input vector I_t in the time step t is multiplied with the weight matrix W^I by eq. (3).

$$I_t W^I = B_t \tag{3}$$

The input vector I_{t-1} in the last time step $t-1$ is multiplied with the weight matrix W^I by eq. (4).

$$I_{t-1} W^I = B_{t-1} \tag{4}$$

Thus the multiplication results B_t in the time step t can be calculated by the results B_{t-1} in the time step $t-1$ and the matrix-vector products $(I_t - I_{t-1})W^I$ by eq. (5).

$$B_t = B_{t-1} + (I_t - I_{t-1})W^I \tag{5}$$

The similarity between the inputs of adjacent time steps largely determines the number of zeros in vector $I_{t-1} - I_t$, and further affects the computation load by the zero-skipping logic in hardware. In other words, the input nodes that have the same values as corresponding ones in the $t-1$ time step can reuse the MAC results B_{t-1}, and need not to be multiplied with the weight matrix in the time step t.

Despite the high degree of similarity between adjacent heartbeats, floating-point network inputs are not exactly the same in two successive time steps in the vast majority of the cases. In order to improve the similarity between successive input nodes in the BLSTM, numerical quantization is adopted. After applying quantization, the BLSTM inputs with close values share the same fixed-point quantitative values, which dramatically increases the degree of similarity. A novel adaptive-grained quantization method is proposed in our design for BLSTM inputs. We use fine-grained quantitative

TABLE 6. Performance comparison of networks with different input quantization strategies.

Method	VEB						SVEB						Similarity
	Acc	Sen	Spe	Ppr	F1	G	Acc	Sen	Spe	Ppr	F1	G	
No Quan	98.6	91.7	99.4	94.0	92.9	92.9	98.3	**80.5**	99.1	82.4	**81.4**	**81.5**	0
Uniform Quan 16	98.5	92.1	99.1	92.1	92.1	92.1	98.0	74.4	99.2	82.3	78.1	78.2	55.6%
Proposed Quan 16	98.5	**92.8**	99.2	92.2	92.5	92.5	98.0	77.0	99.0	79.5	78.3	78.3	**88.7%**
Uniform Quan 32	98.5	92.1	99.1	92.0	92.1	92.1	98.2	76.6	99.3	83.6	80.0	80.0	43.1%
Proposed Quan 32	**98.8**	92.1	**99.5**	**95.5**	**93.8**	**93.8**	**98.3**	78.4	**99.3**	**84.4**	81.3	81.4	49.1%

TABLE 7. Performance comparison of networks with different hidden quantization strategies.

Method	VEB						SVEB						Similarity
	Acc	Sen	Spe	Ppr	F1	G	Acc	Sen	Spe	Ppr	F1	G	
No Hidden Quan	98.8	92.1	99.5	95.5	93.8	93.8	98.3	78.4	99.3	84.4	81.3	81.4	-
Hidden Quan 32	98.8	92.0	99.5	**95.6**	93.8	93.8	98.3	**78.9**	99.2	83.7	81.2	81.2	58.6%
Hidden Quan 16	**98.8**	**92.1**	**99.5**	95.5	**93.8**	**93.8**	**98.3**	78.7	99.3	84.4	**81.4**	**81.5**	74.6%
Hidden Quan 8	98.8	91.6	99.5	95.3	93.4	93.5	98.3	76.7	**99.4**	**85.7**	81.0	81.1	**84.6%**

values on P, Q, S, and T waves, for the details of them contain the majority of information to distinguish heartbeats. Coarse-grained quantitative values are adopted on the R waves, since it not the absolute values but the locations of R waves are of great importance. When the ECG signals are normalized, the R waves usually own higher values and P, Q, S, T waves take place lower values. Thus, the adaptive-grained quantization can be simply achieved according to the signal values and needs not to detect the P, Q, R, S, and T waves. The Table 6 lists a detailed comparison of performance among networks without quantization, with uniform quantization, and with the proposed adaptive-grained quantization on BLSTM inputs, respectively. The results show that the proposed adaptive-grained method with 32 quantitative values achieves a high similarity of 49.1% with negligible accuracy loss owing to the error tolerant ability of neural networks.

The consecutive similar inputs of BLSTM exhibit a high degree of similarity for hidden layers as well. As the distribution of hidden nodes does not have the definite meaning, the common uniform quantization is applied. Table 7 suggests that 16 quantitative values are the best choice with a similarity of 74.6% (the 8 quantitative values result in the decrease of accuracy).

In the proposed BLSTM model with N time steps, the MAC computations between layer nodes (including both input and hidden nodes) and weight matrices in the last N-1 time steps can partly reuse the computing results of the previous time step according to the similarity. Once the corresponding layers nodes are the same in adjacent time steps, the related computations are skipped and the results of previous step are token as those in the current time step. As the similarity of input nodes and hidden nodes are 49.1% and 74.6%, respectively, more than half of MAC operations in BLSTM model can be removed with the utilization of computation reuse.

The computation reuse based hardware architecture for the BLSTM is proposed to take advantage of the similarity between both the input and hidden layers of adjacent time steps. As described in Fig. 5, the coupled layer buffers store the quantized layer nodes at both the last and the current time steps for detecting the similarity of them. The layer nodes at the current time step are subtracted by those at the last time step, and the results are sent to the zero detect model. Only the non-zero remainders and their associated weights are transferred to the PE array for executing the matrix-vector multiplications. Due to the high similarity of consecutive time steps, zero values take place a considerable part of subtraction results. Since the computing results can be reused in the process after, corresponding nodes with the same quantitative values in consecutive time steps only need to be multiplied with the weight matrix once. The Ping-Pong SRAMs are particularly designed for reusing the matrix-vector MAC results between adjacent time steps. The results of the PE array are accumulated in the buffer register and added with the results of last time step from SRAM A to generate final results of current time step, which are written into SRAM B. At the next time step, the roles of SRAM A and SRAM B are exchanged to finish a round of iteration.

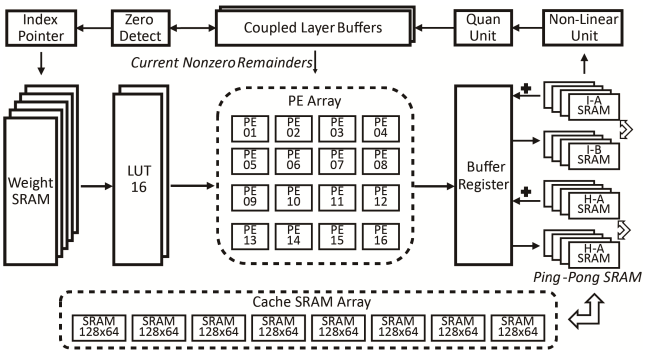


FIGURE 5. The computation reuse based architecture for the BLSTM. The 4-bit quantized weight indexes are read out from the SRAMs and mapped to 8-bit fixed-point format through the LUT. The PE array achieves the multiplication between the remainders of adjacent layers and corresponding weights. The intermediate results are processed by the buffer register files and Ping-Pong SRAM. Zero-detect logic finds the zero related operations to be skipped during the computation. The intermediate MAC results of each time step are stored in the cache SRAM array to achieve the computation reuse design for input overlap.

TABLE 8. Performance comparison of networks with different parameter precisions and BLSTM weight quantization strategies.

Method	VEB						SVEB					
	Acc	Sen	Spe	Ppr	F1	G	Acc	Sen	Spe	Ppr	F1	G
No Quan and Fixed	98.8	92.1	99.5	95.5	**93.8**	**93.8**	98.3	78.7	99.3	84.4	**81.4**	**81.5**
Quan 16 & Fixed 8	**98.8**	**92.1**	99.5	95.1	93.6	93.6	**98.3**	78.9	99.2	83.8	81.3	81.3
Quan 8 & Fixed 8	98.7	91.0	**99.6**	**95.7**	93.3	93.3	98.3	75.8	**99.4**	**85.3**	80.3	80.4
Quan 16 & Fixed 4	95.7	90.5	96.3	72.3	80.3	80.9	97.6	**81.4**	98.4	71.7	76.3	76.4

The non-linear unit achieves the sigmoid and hyperbolic tangent operations for calculating the hidden layer nodes at the current time step before quantization. The the sigmoid and hyperbolic tangent are approximately calculated using piecewise linear function by eq. (6) and eq. (7). The domain of function x in eq. (6) and eq. (7) are divided into multiple pieces, and the actual results of $\sigma(x)$ and $\tanh(x)$ are calculated according to the linear function $f(x) = k_j x + b_j$ in each piece j. The piecewise linear method changes non-linear function into MAC operation with fixed parameters k_j and b_j, which largely reduces the hardware cost in the proposed design. Since the calculation of non-linear functions takes place a small part of the whole BLSTM model, the results of them are not reused among network processing.

$$\sigma(x) \approx \frac{\sigma(b)-\sigma(a)}{b-a} + \sigma(a) = k_j^s x + b_j^s, \quad x \in (a, b] \tag{6}$$

$$\tanh(x) \approx \frac{\tanh(b)-\tanh(a)}{b-a} + \tanh(a) = k_j^t x + b_j^t, \quad x \in (a, b] \tag{7}$$

B. COMPUTATION REUSE DESIGN FOR INPUT OVERLAP

In our BLSTM network topology with N time steps, the inputs of the last N-1 time steps in the current inference is just the inputs of the first N-1 time steps in the next inference. As shown in Fig. 6, since the inputs of each time step are multiplied with the same weight matrix, the input sequence overlap between adjacent inferences allows reusing the results of the current detection in the next detections. As the hidden layer output sequences between adjacent inferences do not overlap with each other, the hidden nodes related MAC operations can not be optimized by this way.

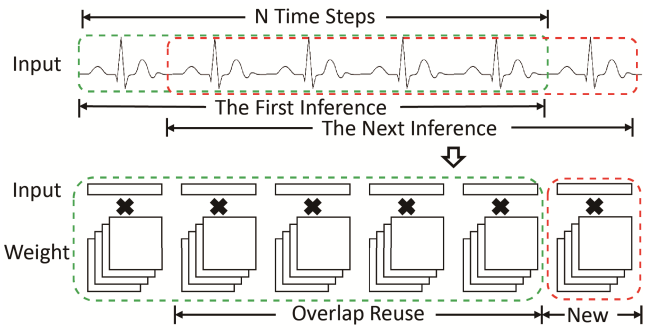


FIGURE 6. The input overlap is utilized to achieve computation reuse for the BLSTM topology. The intermediate results of the last N-1 time steps in the current inference are the same as those of the first N-1 time steps in the next inference. Only the last one time step needs to calculate the new matrix-vector results except for the first inference.

A customized cache SRAM array is used to store the intermediate results, as shown in Fig. 5. When the MAC results of input nodes and weight matrix are generated at one time step, they are stored in corresponding SRAMs and read out as part of the calculating results in the next round of detection. The input overlap based computation reuse results in the 8 kB external memory storage, while reducing 8/9 of the input nodes related MAC operations.

C. NETWORK COMPRESSION DESIGN

The network compression techniques, which mainly include parameter precision reduction and weight quantization, are applied to our design to further minimize the hardware resource and energy consumption.

The intermediate results of MACs are represented in 16-bit fixed-point format to keep the low bits of multiplications, which helps to maintain the acceptable precision of accumulations. When finishing the MAC computations in each time step, the final results are rounded into 8-bit fixed-point format for storage to save hardware resources in the proposed processor. The inputs and weights of all network topologies (including BLSTM, CNN and MLP layers) lower their precision to 8-bit fixed-point format as well. The decrease of precision largely reduces the computation complexity and parameter storage for inference when compared with the original floating-point format.

Since the weights of the BLSTM topology take place the majority of chip storage, they are quantized into 16 8-bit fixed-pointed values in our design. The weights are stored in SRAM with the 4-bit quantitative indexes, which reduces half of the memory size compared with adopting the 8-bit fixed point values. A 16-entry look-up table (LUT) accomplishes the mapping between the quantized 8-bit fixed-point weights and the 4-bit encoded quantitative indexes.

As Table 8 shows, the 8-bit fixed-point parameter format and 16-value quantization operation result in negligible effect on the classification accuracy.

IV. IMPLEMENTATION

A. HARDWARE EVALUATION

The ECG classification processor is implemented in Verilog HDL to evaluate its area, performance and power consumption. The register transfer level (RTL)-level design is synthesized with the Synopsys design compiler (DC) in the SMIC 40LL high-Vt process, while the Synopsys IC compiler (ICC) is applied for placing and routing. The Prime-Time PX is used to analyze the power consumption.

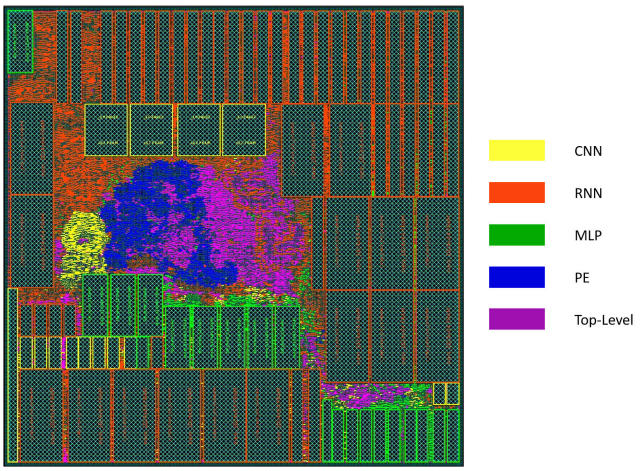


FIGURE 7. Layout of the proposed ECG processor in the SMIC 40LL process.

TABLE 9. ECG precessor implementation data.

Process	SMIC 40nm LL
P&R Area	1.40 mm^2
Voltage	1.1 V
Frequency	100 MHz
Power	2.13 mW
Performance	7.49 GOPS
Energy Efficiency	3.52 GOPS/mW

The layout of the processor is shown in Fig. 7, which has 1.40 mm^2 of total chip area and 186.2 kB of on-chip memory. The details of the layout are listed in Table 9. When simulated under 1.1 V supply and at 100 MHz frequency, the processor consumes 2.13 mW of power. The average processing time per inference is 1.3 ms, translating into 2.78 μJ per beat classification. The execution time of BLSTM, CNN and MLP are 1.18 ms, 0.08 ms, 0.05 ms and the average power of them are 2.13 mW, 3.25 mW, 1.66 mW, respectively. The average energy efficiency achieves 3.52 GOPS/mW, which is state-of-the-art among the existing neural network based processors [22]–[24].

Table 10 compares the proposed design and the prior arts. Compared to the general purpose accelerators for large scale neural network models with compression [22], [23], our ultra-low-power and lightweight processor is more suitable for long battery-life wearable monitoring devices. It is worth mentioning that, our design scale is similar with the CHIP-MUNK [24], whereas our energy efficiency is approximately 3.2x higher and the power consumption is only 7.3% owing to the novel computation reuse based architecture.

B. HARDWARE ARCHITECTURE COMPARISON

To validate the gain obtained through the computation reuse based design, a baseline processor without adopting the proposed hardware architecture is also implemented. Compared with the baseline, our optimized design requires approximately 10 kB larger SRAM capacity due to the extra storage for computation reuse purpose, which results in an increment of average power from 2.069 mW to 2.129 mW at 100 MHz frequency. However, the average number of cycles per each

TABLE 10. Comparison with other neural network processors.

	Proposed	Han [22]	Wang [23]	Conti [24]
Type	ASIC	ASIC	ASIC	ASIC
Technology	40 nm	45 nm	90 nm	65 nm
Platform	Simulated	Simulated	Simulated	Fabricated
MACs	16	64	-	96
Area[mm^2]	1.40	40.8	30.8	0.93
SRAM[KB]	186.2	8384	518.5	82
Frequency[MHz]	100	800	600	168
Power[mW]	2.13	590	1010	29.03
Performance[GOPS]	7.49	510	2460	32.3
Efficiency[GOPS/mW]	3.52	0.86	2.44	1.11

TABLE 11. Comparison with regular processor design.

	Regular Design		Proposed Design	
	10 MHz	100 MHz	10 MHz	100 MHz
Inference Cycle	346826	346826	130533	130533
Computation[MOPs]	9.78	9.78	9.78	9.78
Power[mW]	1.237	2.069	1.289	2.129
Efficiency[GOPS/mW]	0.23	1.36	0.58	**3.52**
Improvement	-	-	2.52x	**2.59x**

detection is significantly reduced from 346826 to 130533. The acceleration of inference makes the proposed design 2.59x more energy efficient than the baseline processor. The detailed comparison is shown in Table 11.

V. CONCLUSION

This paper presents our exploratory work on neural network based processor for multiclass heartbeat classification. Both algorithm and hardware architecture optimizations are employed to avoid excessive resource usage by the neural network methods. The classification algorithm with integrated BLSTM and CNN achieves high accuracy on the arrhythmia detection with a very lightweight network scale. The computation reuse based hardware architecture dramatically accelerates the inference process and significantly reduces energy consumption. Our processor proves to be very efficient for long-term wearable ECG monitoring devices.

REFERENCES

[1] A. B. de Luna, P. Coumel, and J. F. Leclercq, "Ambulatory sudden cardiac death: Mechanisms of production of fatal arrhythmia on the basis of data from 157 cases," *Amer. Heart J.*, vol. 117, no. 1, pp. 151–159, 1989.

[2] N. Bayasi, T. Tekeste, H. Saleh, B. Mohammad, A. Khandoker, and M. Ismail, "Low-power ECG-based processor for predicting ventricular arrhythmia," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 5, pp. 1962–1974, May 2016.

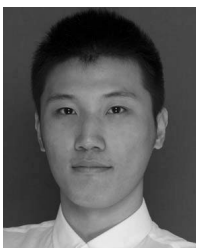
[3] C.-I. Ieong, M. Li, M.-K. Law, P.-I. Mak, M. I. Vai, and R. P. Martins, "A 0.45 V 147–375 nW ECG compression processor with wavelet shrinkage and adaptive temporal decimation architectures," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 4, pp. 1307–1319, Apr. 2017.

[4] Z. Chen, J. Luo, K. Lin, J. Wu, T. Zhu, X. Xiang, and J. Meng, "An energy-efficient ecg processor with weak-strong hybrid classifier for arrhythmia detection," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 7, pp. 948–952, Jul. 2018.

[5] Y. Xu, Z. Chen, F. Li, and J. Meng, "A granular resampling method and adaptive speculative mechanism-based energy-efficient architecture for multiclass heartbeat classification," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 11, pp. 2172–2176, Nov. 2018.

[6] ECAR, AAMI, "Recommended practice for testing and reporting performance results of ventricular arrhythmia detection algorithms," *Assoc. Advancement Med. Instrum.*, p. 69, 1987.

[7] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-based ECG classification for continuous monitoring on personal wearable devices," *IEEE J. Biomed. Health Inform.*, to be published.
[8] S. S. Xu, M.-W. Mak, and C.-C. Cheung, "Towards end-to-end ECG classification with raw signal extraction and deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1574–1584, Jul. 2019.
[9] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," Jul. 2017, *arXiv:1707.01836*. [Online]. Available: https://arxiv.org/abs/1707.01836
[10] S. Chauhan and L. Vig, "Anomaly detection in ECG time signals via deep long short-term memory networks," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2015, pp. 1–7.
[11] M. M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Inf. Sci.*, vol. 345, pp. 340–354, Jun. 2016.
[12] X. Zhai and C. Tin, "Automated ECG classification using dual heartbeat coupling based on convolutional neural network," *IEEE Access*, vol. 6, pp. 27465–27472, 2018.
[13] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.
[14] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[15] A. A. R. Bsoul, S.-Y. Ji, K. Ward, and K. Najarian, "Detection of P, QRS, and T components of ECG using wavelet transformation," in *Proc. ICME Int. Conf. Complex Med. Eng.*, Apr. 2009, pp. 1–6.
[16] P. Warrick and M. N. Homsi, "Cardiac arrhythmia detection from ECG combining convolutional and long short-term memory networks," in *Proc. Comput. Cardiol. (CinC)*, Sep. 2017, pp. 1–4.
[17] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ecg signal classification," *Comput. Biol. Med.*, vol. 96, pp. 189–202, May 2018.
[18] M. Blanco-Velasco, B. Weng, and K. E. Barner, "ECG signal denoising and baseline wander correction based on the empirical mode decomposition," *Comput. Biol. Med.*, vol. 38, no. 1, pp. 1–13, Jan. 2008.
[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[20] R. Mark and G. Moody, *Mit-Bih Arrhythmia Database Directory*. Cambridge, MA, USA: Cambridge, 1988.
[21] A. Amirshahi and M. Hashemi, "Ecg classification algorithm based on STDP and R-STDP neural networks for real-time monitoring on ultra low-power personal wearable devices," *IEEE Trans. Biomed. Circuits Syst.*, to be published.
[22] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, pp. 243–254.
[23] Z. Wang, J. Lin, and Z. Wang, "Accelerating recurrent neural networks: A memory-efficient approach," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 10, pp. 2763–2775, Oct. 2017.
[24] F. Conti, L. Cavigelli, G. Paulin, I. Susmelj, and L. Benini, "Chipmunk: A systolically scalable 0.9 mm^2, 3.08Gop/s/mW @ 1.2 mW accelerator for near-sensor recurrent neural network inference," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2018, pp. 1–4.



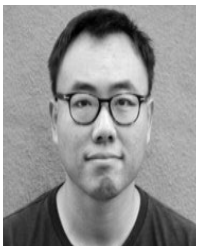
JIAQUAN WU received the B.E. degree in electronic and information engineering from Zhejiang University, in 2015, where he is currently pursuing the Ph.D. degree with the Institute of VLSI Design. His current research interests include biomedical signal processing and neural network accelerating.

FEITENG LI was born in Shijiazhuang, Hebei, China, in 1992. He received the B.S. degree in the Internet of things engineering from the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China, in 2014. He is currently pursuing the Ph.D. degree with the College of Electrical Engineering, Zhejiang University, Hangzhou, China.

His current research interests include physiological signals processing with machine learning and ultra-lower-power neural network accelerator.

ZHIJIAN CHEN received the B.S. and Ph.D. degrees from the College of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2006 and 2011, respectively.

From 2011 to 2013, he was a Postdoctoral Researcher with the College of Electrical Engineering, Zhejiang University. Since 2013, he has been a Lecturer with the College of Information Science and Electronic Engineering, Zhejiang University. His current research interest includes ultra-low-power physiological signal processor design.



YU PU received the B.S. degree from Zhejiang University, Hangzhou, China, in 2004, and the Ph.D. degree in electrical engineering from the Eindhoven University of Technology, The Netherlands, in association with the NXP Research, in 2009. From 2009 to 2011, he was a Research Assistant Professor with Sakurai Laboratory, The University of Tokyo, Japan. From 2011 to 2012, he was a Research Scientist with the Accelerator Team, IBM Research Zurich, Switzerland. From 2012 to 2013, he was a Principal Scientist with NXP Research, where he led Research and Development in ultra-low-power MCUs. From 2014 to 2019, he was with Qualcomm Research, San Diego, CA, USA, and led Research and Development in always-on Android wearable SoC from concept to mass production. Since 2019, he has been with the Computing Research Laboratory, Alibaba DAMO Academy, Sunnyvale, CA, USA. He has authored or coauthored over 30 scientific publications and holds more than 15 US patents. Dr. Pu is the Technical Program Committee (TPC) Chair and a member of various IEEE/ACM conferences. He is currently an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS (TCAS-I).

MENGYUAN ZHAN received the B.S. degree from Weifang Medical University, Weifang, China, in 2013. She is currently with the Affiliated Hospital of Qingdao University, Qingdao, China.