Energy-Efficient Arrhythmia Classifier Using Layer-Wise Quantized Convolutional Neural Network

*Abstract*—Arrhythmia diagnosis neural network can perform continuously monitoring, real-time diagnosis and potential risk waning in wearable devices. However, the existing neural networks suffer from high memory and power consumption which limit the application of the diagnosis neural network in low-power wearable devices. Here, we proposed a novel neural network approach to classify 17 rhythm classes using 1,000 long-duration electrocardiograms (ECG), achieving a classification accuracy of 95.72% that is 4.32% higher than current state-of-the-art methods. And we proposed a layer-wise quantization method based on The Greedy Algorithm and perform comparable to other quantization methods. With the combination of the two works, we achieve 95.39% classification accuracy and reduce the memory consumption by 15.5 times. Our study leverages low-memory neural network for high performance and low power consumption, and demonstrates the possibility of implementing neural network in wearable devices for continuously monitoring, real-time diagnosis and potential risk waning.

*Index Term—*Neural Network, Arrhythmia, Quantization, The Greedy Algorithm, Energy-Efficient, Wearable Device.

Ⅰ. INTRODUCTION

Arrhythmia is the important cause and manifestation of cardiovascular diseases (CVDs) morbidity and mortality[1], which are the leading cause of mortality worldwide[2]. The ECG analyzation is one of the most common methods to diagnose arrhythmias, and it’s usually done in hospitals with experiment cardiologist. However, early symptoms of some arrhythmia may be hard felt due to their short duration, some symptoms are sudden and intense in several minutes[3], which can lead to serious consequences without timely treatment like stroke, heart failure, coronary artery disease and even life-threatening­[4]. The time-consuming and location-specific testing method may limits the patients or potential patients monitoring their daily cardiac activity and might delay timely treatment. To tackle this issue, the low-power automatic arrhythmia diagnosis wearable devices which provide continuously monitoring, real-time diagnosis and warning the potential risk have been a research hotspot in both biomedical and computer[5].

Neural network is the essential part of driving the substantial algorithmic advances in recent years[16]. The hierarchical structure of neural network have made it learn abstract and intrinsic features and shown better generalizability and robustness in computer vision and natural language processing. Compared feature select-based algorithm, neural network learns features from the raw ECG waveforms without requiring extensive feature designing or preprocessing makes it a particular simple algorithm for arrhythmia diagnosis. Therefore, the neural network-based arrhythmia diagnosis algorithm attracts more and more attention.

Nevertheless, neural network methods are always associated with high resource consumption, which is paradoxical to the requirement for deploying the arrhythmia classification neural network to the resource-limited hardware platform, because of its high computing complexity and large memory requirement. In other words, solving the power consumption problem of arrhythmia classification will push the enormous advance of automatic arrhythmia diagnosis.

In order to address this limitation, we propose a novel neural network for arrhythmia classification and a novel layer-wise quantization method based on Greedy Algorithm for compressing the memory and power consumption without excessive accuracy loss. The contributions of the paper are threefold:

1. Firstly, a novel long-duration ECG fragments CNN architecture for classifying arrhythmia is proposed, and achieves the arrhythmia classification of 95.7% for long-duration ECG classifying;
2. Secondly, a Greedy Algorithm-based layer-wise quantization method is proposed to lower the memory and power consumption of the neural network, while maintaining the high classification ability.
3. Finally, we implement the quantization method to the neural network and achieve the classification accuracy of 95.39% with compressing the memory by 15.5 times, which is better than other works.

The rest of this paper is organized as follows: In Section II, the related work is reviewed. Section III introduces the proposed convolutional neural network architecture and layer-wise quantization scheme. Section IV gives the experimental results and discussion. Finally, conclusions are given in Section V.

Ⅱ. RELATED WORK

Most existing automatic arrhythmia diagnosis methods use pattern recognition. These methods perform hand-crafted feature extraction by transforming the input signals into a variety of features with cardiological knowledge, then sent the features to a followed classifier for arrhythmia diagnosis. Up to date, the massive amount of different features has been proposed. For example, [6] determines the heartbeat fiducial points manually to obtain futures related to heartbeat, calculates features from RR-intervals and heartbeat segmentation information, and uses two linear discriminant classifiers (LD) for the final classification. However, the morphology features are often difficult to detect because of their low amplitudes, incurring bad results. Thus, [7] proposed a hierarchical classification method based on [6]. It applied higher-order statistics (HOS) and Hermite basis functions (HBF) to extract similarities features and feed the features to a hierarchical classification method. Different from the former systems using heartbeat, [9] use long duration ECG signal for arrhythmia classification. Welch’s method and a discrete Fourier transform are conducted to estimate the spectral power density of the ECG fragments, and a Support Vector Machine achieves high accuracy than other classifiers including K-Nearest-Neighbor, Probabilistic Neural Network, and Radial Basis Function Neural Network. Although the above methods achieve high accuracy, the performance is mainly depends on the discriminative features devised which have poor generalizability for a broad range of patients[10].

Recently, end-to-end arrhythmia neural networks are proposed successively. The method fuses the feature extraction phase and classification phase that the type of heart rhythm is directly output with the raw ECG signal input. For example, [11] used a neural network consisting of a three-layer CNN and two-layer MLP to learn features of different presentation of patient-specific downsampled heartbeat, the results show that the algorithm outperformed any arrhythmia classification algorithm existed at that time in diagnosis of all classes except class S of arrhythmia. Considering the performance of [11] heavily relies on manual annotations of input data and the network is patient-specific, [12] used Pan-Tompkins algorithm [13] to segment heartbeat and a 5-layer DNN for arrhythmia diagnosis. This patient-independent ECG classifier proposed by [12] achieves a 10% higher sensitivity at least than state-of-the-art methods in detecting supraventricular- and ventricular-ectopic beats. As only a small proportion of abnormal heartbeat in cardiac activity, it’s ineffective to accurately detect all beats. [14] built a two-stage neural network that the first stage MLP classifies only the normal and abnormal heartbeat and the second one is used for further diagnosis of abnormal heartbeat. In the meanwhile, an adaptive ECG compression method is offered to further energy reduction. The aforementioned heartbeat neural networks are limited since the input sequences acquired beat detection to divided into single heartbeat as fixed-length input of neural networks. Analysis long-duration ECG signals eliminate heartbeat detection. [15] used a 1D-CNN diagnosis a 10s ECG signal, achieving a truly end-to-end diagnosis. And [16] developed a neural network with 1.28s signal as input. Comparing with Cardiologists, it performs high diagnostic performance similar to that of cardiologists.

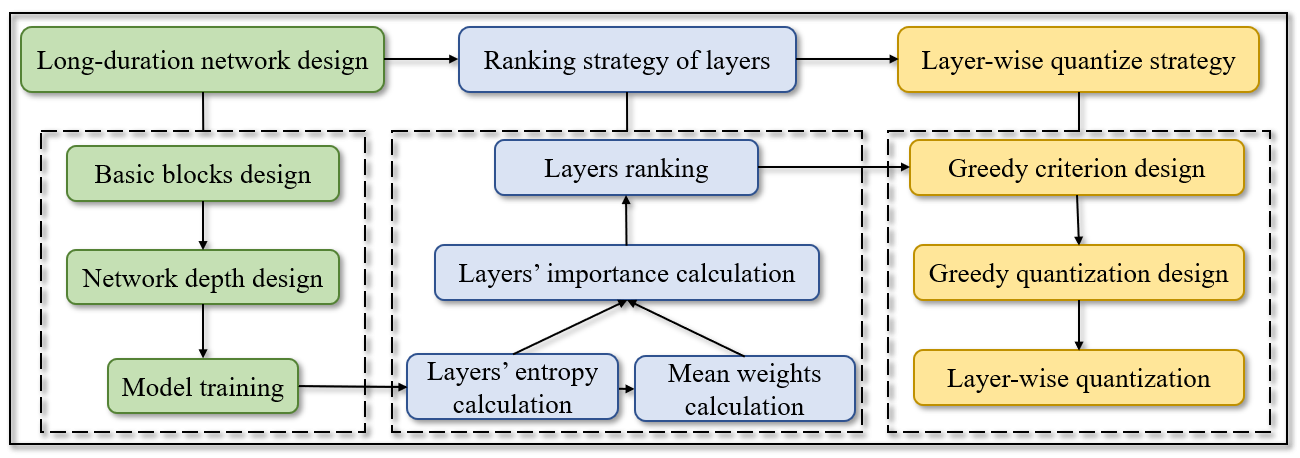
However, as these method involve neural networks, they’re both computationally intensive and memory intensive, making it difficult to implement on embedded systems with limited hardware resources.

And with the evolution of neural network for arrhythmia diagnosis, neural network quantization provides new ideas which can lower the memory and power consumption. Quantization is the process of constraining weights from high bit to low-bit ones with minimal performance degradation. According to [17], a 2.77 GB ECG signal is generated for real-time monitoring per day that requires a massive memory and energy of data transmission and processing. [20] has shown that neural networks can inference with 8-bit weights. [21] constrains the weights into +1 or -1 during inference that get rid of the multiply-accumulate operations and achieve nearly state-of-the-art results. To expand the presentation of the quantized neural network, [22] quantized all weights and activations to a fixed bit in both training and inference, and the neural network can be implemented on embedding systems with the quantized weights. [23] proposed an adaptive layer-wise quantization method for more flexible quantization configuration, it applies entropy of weights to define quantization bit for each layer. It’s worth noting that these methods could combine with neural networks for arrhythmia diagnosis considering the similarity of structure and mechanism, despite the above quantization methods are aimed at the neural networks for image classification.

Ⅲ. METHODOLOGY

In this section, we first introduce the motivation to build a hardware-oriented arrhythmia classifier. Then, the system overview of the classifier is provided, and next, we describe the detail of the convolutional neural network. The quantization strategy and layer-ranking procedures in the quantization strategy are discussed at the end of this section.

The overall framework of the arrhythmia detection system proposed is shown in Fig. X, which can be divided into three stages. The first stage is the design of arrhythmia detection neural network for long-duration ECG signals, which is based on the basic blocks design and the exploration for the depth of neural network to achieve the best classification accuracy. The second stage is the ranking strategy of layers, which ranks layers of the designed neural network through the integration of weights entropy and mean of each layer. And the third stage is layer-wise quantization based on the Greedy Algorithm, selecting a bitwidth of weights for each layer to achieve better classification accuracy and lower memory requirement.

Fig. X. The flowchart of the proposed framework

1. *Design of Convolutional Neural Network Using Long-duration ECG Fragments*

Initially, we developed a novel arrhythmia classification convolutional neural network shown in Fig. X. The end-to-end network receives the raw and long-duration ECG fragments as input and outputs the arrhythmia class to which the input belongs without any data processing.

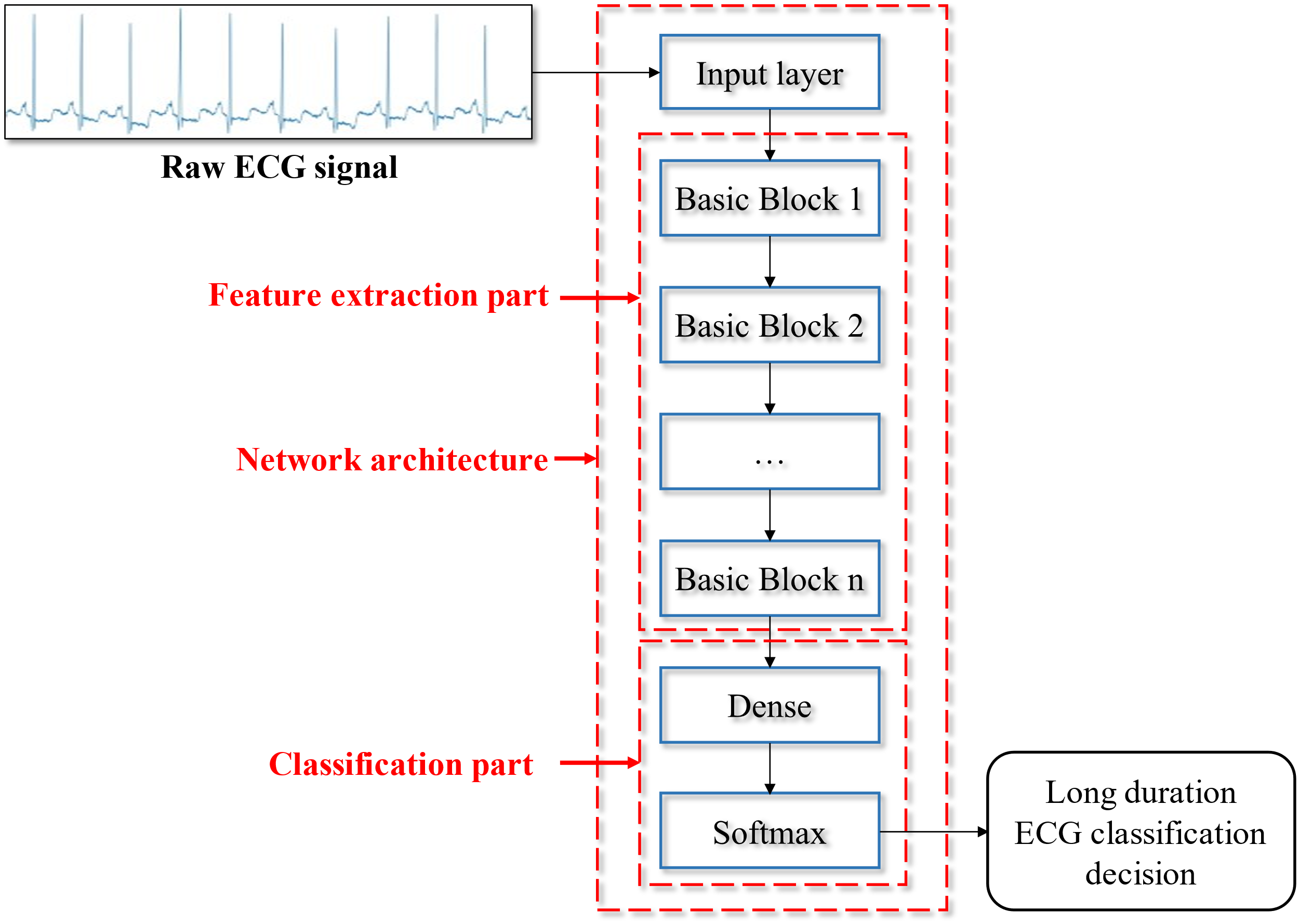


Fig. X. End-to-End arrhythmia classification neural network using long-duration ECG fragments.

The proposed neural network has a cascade architecture which is similar to LeNet[24], as illustrated in Figure X. In detail, the network is composed of several basic blocks and two fully connected layers. The basic block consisting of a 1D-convolution layer and a 1D-maxpooling layer is demonstrated in Fig. X. They are utilized to extract features from the long-duration ECG signals. And the feature extraction ability of each basic block is determined by two parameters: kernel size *k* and filters *m*. Considering that the morphological characteristics of ECG waveforms are informational features for the diagnosis of arrhythmia, the proposed neural network employs larger convolution kernels to extract features contained in a broad range of sampling points to obtain the information contained in the ECG during different cardiac activities. So the kernel size *k* and filters *m* are limited to 3-13 and 4-256 respectively. Behind the stacked basic blocks, two fully connected layers are used to arrhythmia classification with the learned features. The former one is used to convert the 2D-feature maps to a 1D series and learn the global feature, which has 64 hidden neurons. And the latter one is a softmax layer with 17 neurons of which each output is considering as probability to the corresponding heart rhythm. The softmax function is as followed, where  is the number of the rhythm type:





Fig. X. A basic block

The input of our network is a 10s ECG fragment (sampled in 360Hz), and end-to-end outputs the arrhythmia class to which the input belongs without any data processing. Compared with the heartbeat-based arrhythmia classification networks[12], our network requiring no QRS complexity detection which can not only lower the power consumption to the hardware implementation but also improve the algorithm performance. That is the input of a neural network is usually a fixed-length sequence, but the duration of different heart rhythms is different, for the purpose of getting a complete heartbeat, QRS complexity detection is unavoidable in the heartbeat-based neural network before entering the network. Moreover, the division between heartbeats might lead to performance degradation, due to the information loss in adjacent heartbeats which is sufficiently important for arrhythmia.

Besides, some studies demonstrate that the depth of neural networks may influence the network performance[25], but a deeper architecture will bring additional power consumption and memory requirement. Loosely inspired by Moore[26], we apply a brute investigation to search the optimal number of basic blocks used in the neural network which has the optimum configuration in both classification accuracy and memory consumption. We experimented with 8 depth configurations and compared the performance of each one. The result shows that the neural network with 7 basic blocks achieves the highest classification accuracy of 95.7% while occupying the smallest memory of 314.75KB. And a further experimental result on the neural network depth will be described in Section-Ⅳ.

1. *Layer-Wise Quantization Strategy*

Power consumption is one important factor limits the implement of neural network in arrhythmia automatic diagnostic hardware. To tackle this issue, we proposed a layer-wise quantization method based on the Greedy Algorithm to effectively reduce hardware consumption.

Transmission and computation are the main sources of power consumption in neural network for arrhythmia diagnosis, hence compression is the common method to reduce hardware consumption. ECG signal compression methods are broadly evolved[27], however, rarely of research combine neural network compression with arrhythmia classification. Actually, it is believable that neural network compression might contribute to effectively reduce hardware consumption. That is the working mechanism of neural networks makes weights in neural networks reusable. At the same time, the interaction between layers causes different layers have different effects on network performance.

According to the aforementioned hypotheses, we proposed an approach to search and achieve layer-wise quantization strategies based on the greedy algorithm, and the objective is to find the optimal quantization strategy that can be implemented in arrhythmia automatic diagnostic hardware with minimal energy and minimal performance degradation. And the flowchart details our layer-wise quantization strategy in Fig. X.



Fig. X. Flowchart of the proposed layer-wise quantization

It is notable that the order of layers in layer-wise quantization isn’t their position in the network, rather than ranking according to their weights distribution. A specific description of the ranking method will present in Section-Ⅲ C.

Based on the sorted layers, the floating-point weights are obtained, as well as the number of computations , number of weights , and number of activations of each layer. A symmetric quantization, shown in Algorithm 1, is applied to the sorted layers sequentially. First, we set the maximum absolute value of weight as , and then weights are scaled by to obtain the normalized weights (). And the range of will further divide into discrete values of the corresponding bit width. It should be noted that the true form of weights is used here to represents both positive and negative numbers, so 0 has two representations: +0 and -0, but only one is used, so there are only quantization levels, and will be replaced by the nearest quantization level . Finally, are returned as the quantized weights.

|  |
| --- |
| **Algorithm 1** Symmetric Quantization |
|  |

After getting the quantized weights , the neural network will replace the float-point weights with and recalculate the classification accuracy in the validation set. When all of the accuracies of possible bit widths in this layer are acquired, they are delivered to the Algorithm 2 (Greedy Criterion) along with the number of computations , number of weights , and number of activations of this layer. Greedy criterion will guide the algorithm to select the optimal quantization strategy in both neural network performance and power consumption. Moons’ s[28] energy model is applied to calculate the power consumption of neural network in inference (line3-6), the in line 4, 5 is set to 6 as the reduction in memory-access due to the parallelism in hardware. The total energy per inference is the sum of computing energy and memory-access energy of weights and activation :

The *p*, the parameter of reuse time in hardware, is set as 6.

|  |
| --- |
| **Algorithm 2** Greedy Criterion |
|  |

For the same layer, , and are all fixed values, so of different bit widths of the same layer are only associated with the bit width. However, considering the amount of the three terms vary by layers, if only the is used, Greedy criterion tends to adopt an adaptive quantization strategy that using lower bit widths for high-consumption layers, whereas, high bit widths for low-consumption layers . And it makes the quantization strategy inflexible. Therefore, we innovatively propose using the energy-consumption curve slope instead of , that is, using the least-squares method to fit the power consumption of the same layer at different bit widths and using the slope of the fitted curve instead of for quantization strategy selection. The following function is used to evaluate and select the quantization strategy with highest for this layer, where β is the coefficient for flexibly guiding the selection.

When the optimal bit width of one layer is determined by the Greedy criterion, the weights of this layer will be quantized to this bit width and replace the float-point weights to quantify the next layer until all layers are quantized. At the same time, we also quantify activations of the previous layer, which is important to save energy by computing with the same bit width. After all of the layers quantized, the final quantization strategy for the neural network is determined, and it is possible to adjust the selection by changing the coefficient *β*.

Under our quantization scheme, an efficient flexible and layer-wise quantization strategy is available. Layers of neural networks are interrelated because of the positions, so the optimal solution can only be obtained by brute search method on the quantization strategies of every layer or dynamic programming. However, these two methods are time-consuming and labor-intensive, the optimal solution may even change when weights change. This method can accelerate quantization selection: Take an 8-layer neural network as an example, the weight bit width of each layer parameter ranging 1-8, the brute search needs to be performed  times, while our quantization method demand only 64 times to find the suboptimal strategy and co-design of performance and hardware consumption is also provided

1. *The Ranking Strategy of Layers*

This section will detail our ranking strategy and provide two different strategies. Because we use the layer-wise quantization strategy, the input-output relationships of layers and each layer are interrelated, it is necessary to decide the order of the layers to perform hierarchical quantization. The ranking strategy is an emerging theme of our research because the performance of layer-wise quantization will suffer a huge impact by the ranking strategy.

The view that large weights and activations have a significant impact on the network has been widely accepted[29], but different quantization methods have the opposite treatment for the large weights. [29] show that preserving the large weights as high-precision can reduce quantization error and their frequency brings little hardware consumption. However, [30] shows that the large weights are infrequent and low-precision leads to less performance degradation but maximizes the utility of quantization levels. We adopt the view that large weights have a great impact on neural network performance and consider that the weights’ number of different layers varies greatly, so the mean of the absolute value  of each layer is used as one of the indicators for layers sorting shown in Eq. (X), where  is the index of layer:

 (X)

Besides, entropy  is the other indicator in the work at which information is produced by a stochastic source of data. The larger the entropy, the more information the data contains, and the more the data is confusing, which is widely used in the field of neural network compression. The calculation of entropy is as follows, where  is the probability of the *i*th weight:

 (X)

The Eq. (X) details our formulation of the ranking strategy, where  is the mean of the absolute value of each layer,  is the entropy of each layer and α is the coefficient to balance the two indicators.

 (X)

And layers with higher *I* are thought to have a greater impact on network performance. In other words, quantization for the high-*I* layers might incur a more significant performance drop than the lower. With this opinion, we have proposed two method to layer-wise quantization. The first one is performance-predominant method, which quantizes the high-score layers first and then quantizes the low-score layers. On the premise of minimizing performance error, the high-*I* layers tend to maintain high-precision weights, and the low-*I* layers also use high-precision weights for the high-score layers’ input that leads to this strategy having a conservative compression rate and low-performance error. The second method, memory-predominant method, quantizes the low-score layers first and then quantizes the high-score layers, due to the low-*I* layers can tolerate lower precision of weights without excessive performance reduction and achieve better compression rates. And quantization for layers with higher scores will use conservative bit width to avoid further performance drop.

Ⅳ. EXPERIMENT SETTING

This section firstly introduces the dataset, and presents our evaluation method. Then, it describes the experimental of neural network depth and compares the our best neural network’s performance with current state-of-the-art methods. Next, we detail the layer-wise quantization experiment in the selected neural network topology. Finally, we demonstrate the comparison result among the performance-predominant method, memory-predominant method and other neural network quantization methods.

1. *Dataset*

The long-duration ECG dataset used to neural network training and performance evaluation is the dataset [31]. And all of the ECG fragments in this dataset are from the MIT-BIH Arrhythmia dataset[32] of PhysioNet. The MIT-BIH Arrhythmia contains 48 30-min, two-leads ECG waveforms, and all of the QRS complexes are labeled accordingly. And all of the ECG data are filtered at a band frequency of 0.1-100 Hz and then sampled at 360 Hz. The characteristics of the dataset are: (1) there are 1,000 10-second fragments of data, and samples are non-overlapped with others; (2) the signals are from 45 patients: including 19 females (aged 23-89) and 26 males (aged 32-89); (3) the contains 17 categories: normal sinus rhythm, pacemaker rhythm and 15 types of arrhythmia, each type contains at least 10 segments; (4) all signals are from modified limb lead II. A more detailed distribution of heart rhythm types is shown in Table X:

**Table X**. The description of ECG samples used for the various heart rhythm classes[29]

|  |  |  |
| --- | --- | --- |
| Index | Class Name | No. of Instances |
| 1 | Normal sinus rhythm | 283 |
| 2 | Ventricular tachycardia | 10 |
| 3 | Idioventricular rhythm | 10 |
| 4 | Ventricular flutter | 10 |
| 5 | Fusion of ventricular and normal beat | 11 |
| 6 | Left bundle branch block beat | 103 |
| 7 | Right bundle branch block beat | 62 |
| 8 | Second-degree heart block | 10 |
| 9 | Pacemaker rhythm | 45 |
| 10 | Atrial premature beat | 66 |
| 11 | Atrial flutter | 20 |
| 12 | Atrial fibrillation | 135 |
| 13 | Supraventricular tachyarrhythmia | 13 |
| 14 | Pre-excitation (WPW) | 21 |
| 15 | Premature ventricular contraction | 133 |
| 16 | Ventricular bigeminy | 55 |
| 17 | Ventricular trigeminy | 13 |
|  | Total | 1000 |

In the experiment, 70% of the total samples are randomly selected as the training dataset and the remaining 30% as the test dataset. And both of the datasets are completely disjointed sets of patients. Besides, considering that the extreme data imbalance of different classes in the training phase, we performed an over-sampling to the low-number classes to balance the number of classes.

1. *Evaluation Method*

In this paper, we evaluate both of the performance of the neural network classification and the effectiveness of the layer-wise quantization method separately.

The evaluation methods of neural network classification are the overall accuracy (*OA*), specificity (*Spe*) and sensitivity (*Sen*) of heart rhythms from the input samples. These evaluation function are presented in Eq. X - Eq. X, in where TP is True Positive, FN is False Negative, FP is False Positive and TN is True Negative.









The inference power and compression rate (*CR*) are used to evaluation the effectiveness of neural network quantization method. We adopt the inference power model from [28] to calculation the power consumption of neural network. The inference power model is shown in Eq. X - Eq. X that the total energy per inference is the sum of computing energyand memory-access energy of weights and activation. *CR* is the ratio of memory occupied by the neural network before and after quantization.



1. *Convolutional Neural Network Topologies*

In this experiment, we carried out experiments on the relationship between the neural network depth, classification accuracy and the memory requirement, then determine the final network topology based on the two evaluation methods. Fig. X presents the diagram of 8 types of network topologies with different depths composed of basic blocks. The architecture of all topologies consists of multiple basic blocks and 2 fully connected layers connected in series. The basic blocks are used to extract features from ECG, and the two fully connected layers are used to synthesize global features for classification.



Fig. X**.** diagram of 8 neural networks

We use the same training sample to train the 8 neural network topologies for 3,000 epochs, and then select the highest *OA* during the training process as the classification accuracy. Statistics on the *OA* and memory requirement of networks with different topologies are shown in Fig. X. The memory requirement here is considered as the hardware consumption because it can affect both memory and the energy directly when is implemented on a hardware platform. As seen from Fig. X that the classification accuracy increases with the increase of the number of basic blocks. When the number of basic blocks is 5, the classification accuracy reaches saturation with the *OA* of 94.7%. And when the number is 7, the *OA* reached the highest, which was 95.7%.

As for the memory requirement, it presents a more complex trend with the changing of basic blocks, that is, the increase of basic blocks brings more weights in convolutional layers requiring more memory, while the number of input hidden neurons of full connect layers decreases with the increase of convolutional layers. Therefore, the memory requirement in 1-5 is complicated. And the memory requirement is the lowest when the basic block is 7. In summary, the required memory reaches a local minimum of 343KB and 314.75KB respectively when the number of basic blocks are 2 and 7.

**Fig. X.** Impact of the number of basic blocks on memory consumption and overall accuracy of arrhythmia diagnosis

To sum up, when the number of basic blocks is 7, the *OA* is 95.7% at the highest and the minimum memory requirement of 314.75KB. Therefore, this structure configuration with 7 basic blocks and 2 fully connected layers are used in the subsequent experiments. And its architecture and parameter distribution are shown in Table X.

**Table X**. architecture and parameters distribution of the neural network

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer** | **Conv** | **Pooling** | **Parameters** |
| Basic Block1 | 16×1×8 | 8×1 | 136 |
| Basic Block2 | 12×1×12 | 4×1 | 1164 |
| Basic Block3 | 9×1×32 | 4×1 | 3488 |
| Basic Block4 | 7×1×64 | 4×1 | 14400 |
| Basic Block5 | 5×1×64 | 2×1 | 20544 |
| Basic Block6 | 3×1×64 | 2×1 | 12352 |
| Basic Block7 | 3×1×72 | 2×1 | 13896 |
| FC1 | 216×64 | - | 13888 |
| FC2 | 64×17 | - | 1105 |

1. *Diagnosis Performance and Comparison*

Fig. X shows the confusion matrix of the proposed neural network with 7 basic blocks using the test data. The proposed neural network correctly classifies 290 out of 304 samples during the test phase yielding 95.72% of overall accuracy. In the test phase, 10 of the 17 heart rhythms are classified without misdiagnosis, and the target class is dominant in individual class’s prediction. Moreover, the confusion matrix also demonstrates that 91.7% of the low-number class samples are correctly classified with our over-sampling strategy.



Fig. X. Confusion matrix for the predictions of the CNN versus the targets

Table X compares our neural network with two state-of-the-art long-duration arrhythmia classifiers [15],[33] in the context of classification accuracy and the memory requirement. And all of the work is tested with the dataset[31]. It can be observed from Table X that our neural network outperform than both [33] and [15], the proposed neural network achieves *OA* of 95.72%, improving the overall accuracy by 4.32% and 4.39%. Furthermore, compared with the neural network method[15], our neural network not only improves the *OA* by 4.39% and even reduce the memory requirement by 24.8 times.

**Table X**. The description of ECG samples used for the various heart rhythm classes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Sen*(%) | *Spe*(%) | *OA*(%) | Memory(KB) |
| Yıldırım [15] | 83.91 | 99.41 | 91.33 | 7852 |
| Pławiak[33] | 91.40 | 99.46 | 91.40 | - |
| Proposed | **94.68** | **99.71** | **95.72** | **314.75** |

1. *Compression Performance and Comparison*

In this experiment, we applied our layer-wise quantization method described in Section Ⅲ to the above neural network.

The violin plot in Fig. X shows the statistics on the weights distribution of each layer in the proposed neural network. Each part in violin plot corresponds to the weights distribution and probability density function fitting curve of a single layer. Observing the distribution in Fig. X can be summarized as follows: (1) all weights in this neural network are concentrated in the range of [-0.6, 0.55]; (2) the weights of each layer is roughly symmetrical about 0; (3) weights in most layers present Gaussian distribution except FC2; (4) The weights of each layer is relatively concentrated with few layers having a few outliers.

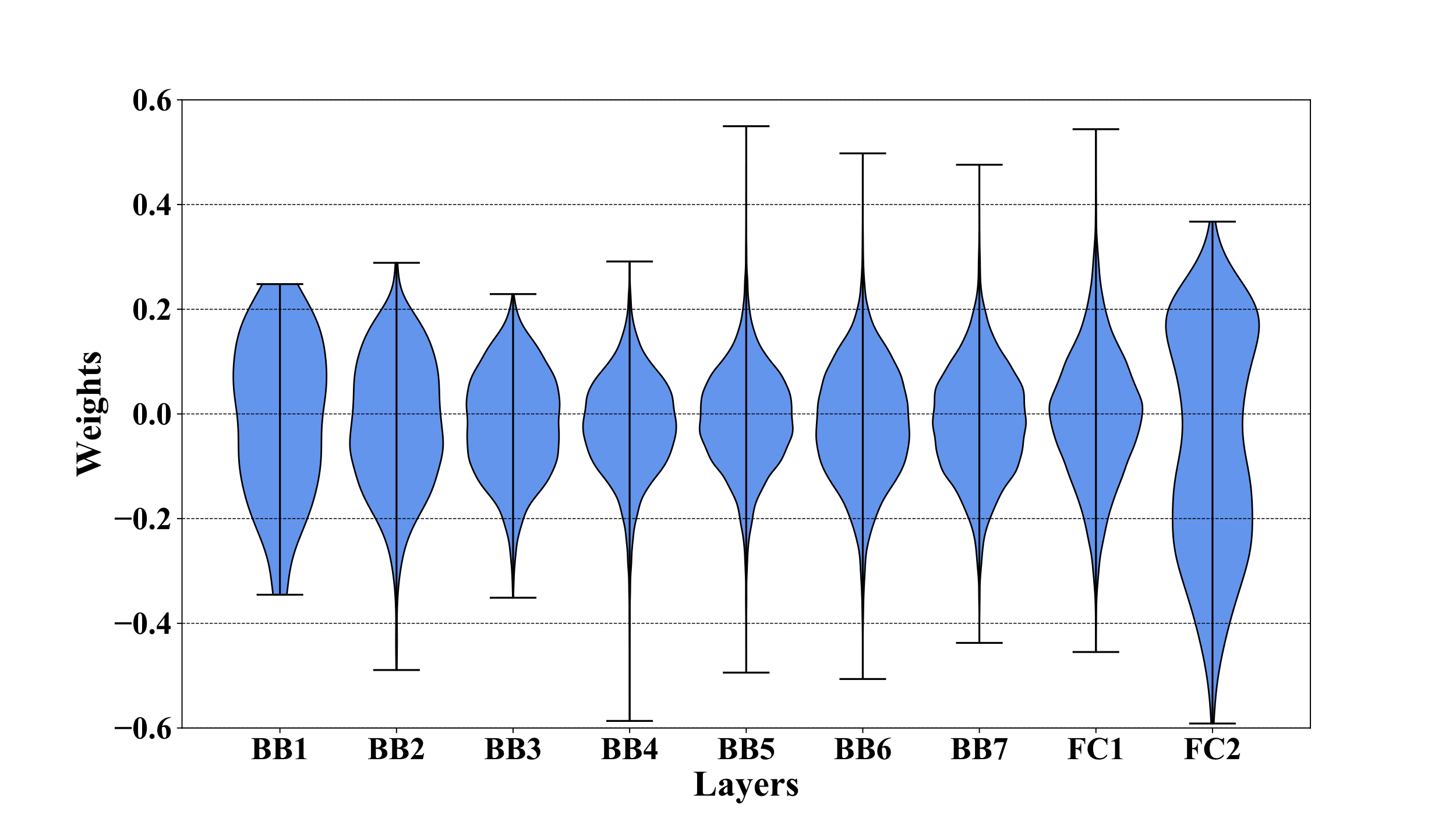


Fig. X. Weight distribution of layers

According to the weight distribution of the neural network, the aforementioned ranking strategy is applied to calculate the mean of the absolute value and information entropy of each layer. The calculation result is shown in Figure X. It can be seen that both of the mean and entropy show a distribution trend that is high near the input and output and low in the middle layers approximately. And the FC2 has the highest mean and entropy, while the Basic Block4 has the lowest.



Fig. X. Entropy and mean absolute weights of layers

Based on the Eq. (X), the *I* of each layers are calculated with the α of 0.5 and shown in Table X. The FC2 has the highest importance and BB4 has the lowest importance. The importance is generally distributed high at both ends of the structure and low in the middle layers.

**Table X**. The Importance order of layers

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ranking | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Layer | FC2 | BB2 | FC1 | BB1 | BB6 | BB7 | B3 | BB5 | BB4 |

According to the above ranking results, we use the 2 above quantization methods for the neural network: the performance-predominant quantization method starts with FC2 with the highest ranking and performs layer-wise quantization, and the memory-predominant quantization starts with BB4 with the lowest ranking score. During the layer-wise quantization process, we recorded the *OA*, memory consumption and energy. And Fig. X shows these change curves under performance-predominant and accuracy-predominant quantization methods.

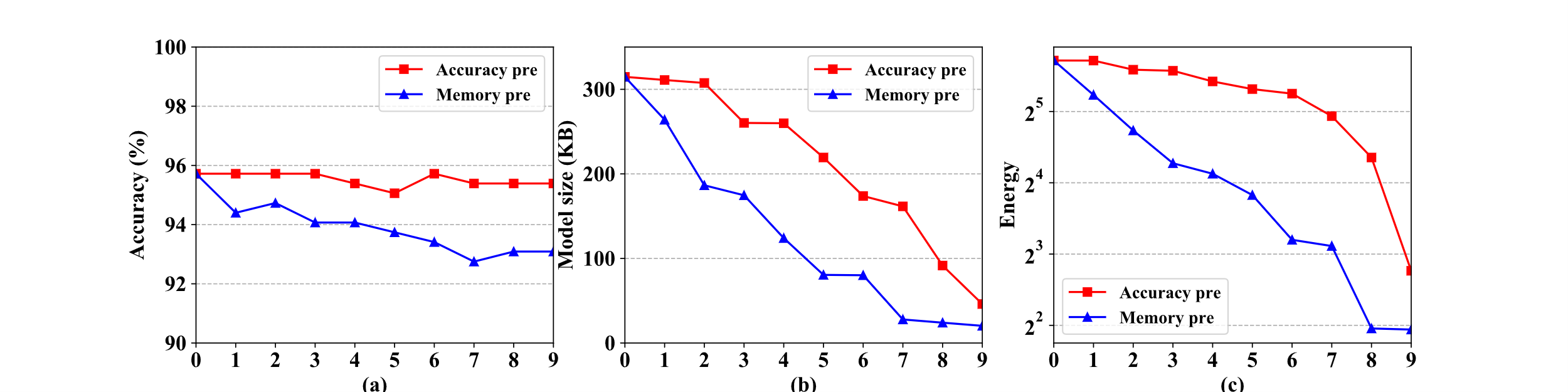


Fig.12. Evaluation during quantization. (a) Overall accuracy, (b) Memory consumption (c) Energy.

Fig. 12a illustrates how the *OA* changes during the proposed layer-wise quantization process. The accuracy-predominated mode has higher accuracy than the memory-predominant mode. And it’s clear that the *OA* remains basically unchanged during the quantization process. Although the *OA* in memory-predominant mode decreases larger, the *OA* is still 93.09% with only 2.63% decreasing than the un-quantized network. In addition, it can be observed that the *OA* has increased in the two modes during the quantization process. That is due to the layered topology of neural network induce interaction among layers, which is one important reason for performing layer ranking and layer-wise quantification. Fig. 12b shows the memory consumption change during the quantization. The memory consumption in accuracy-predominated and memory-predominated modes reduce significantly with the *CR*s of 6.83× and 15.50× respectively. Fig. 12c is the trend of power consumption in the layer-wise quantization process. Accuracy-predominant and memory-predominant reduce the memory capacity by 7.73× and 13.67× respectively. As discussed above, the memory consumption and energy of the memory-predominant mode is less than half of the accuracy-predominant mode, which means that the former one is more suitable for wearable devices with memory and energy restrictions.

The quantized bit width of each layer with the two modes is shown in Table. X. It can be found that the bit width of most layers in accuracy-predominant is higher than the memory-predominant mode, which is consistent with the assumption made in Section Ⅲ. The bit width of most layers after quantization using the memory-predominant mode is relatively low, and some of the layer's bit width is even set to 1 bit, that is, all of weights in these layer are quantized to ± 1 which can reduce memory and energy greatly.

**Table X**. Bit width of layers in accuracy-predominant and memory-predominant mode

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mode | BB1 | BB2 | BB3 | BB4 | BB5 | BB6 | BB7 | FC1 | FC2 |
| Accuracy-pre | 7 | 7 | 3 | 6 | 4 | 5 | 5 | 4 | 4 |
| Memory-pre | 5 | 5 | 4 | .3 | 1 | 6 | 2 | 1 | 4 |

Table X compares the classification accuracy and memory consumption of the proposed quantization method with several state-of-the-art work. For fair comparison, all of the quantization methods are work on aforementioned neural network. As seen from Table X, the accuracy-predominant model achieve the highest *OA* of 95.39 with only 0.33% loss than the unquantized model. And the memory-predominant reduce the memory to half of 20.31KB with a 2.30% than the former that is still higher than other methods. Binary Connect[29] compress all weights to 1 bit and achieve the maximum *CR* of 32, but its *OA* is only 56.25% that is too low to be used in arrhythmia classifier. As for DoReFaNet, weights are quantized to 2 and 3 bit. In 2-bit quantization, the memory is reduced by 16 times, but the *OA* is the lowest. Although the *OA* of DoReFaNet with 3-bit weights improves by 31.78%, it is still 8.22% lower than the memory-predominant with nearly same memory consumption. The comparison indicates that the proposed layer-wise quantization method based on Greedy Algorithm is effective than other state-of-the-art work, and the two quantization modes even allow more flexibility in configuring network compression.

表X 不同量化模式的识别准确率与内存占用性能对比

|  |  |  |
| --- | --- | --- |
| Mode | *OA* (%) | Memory (KB)/Compression rate |
| Binary Connect[29] | 56.25% | **9.84/32.00×** |
| DoReFaNet (2 bit) [33] | 53.09% | 19.67/16.00× |
| DoReFaNet (3 bit) [33] | 84.87% | 29.50/10.67× |
| Accuracy-predominant | **95.39** | 46.07/6.83× |
| Memory-predominant | 93.09 | 20.31/15.50× |

Ⅴ. CONCLUSION

Memory and energy limit the application of neural network in wearable heart rhythm detection devices. In this work, we propose a novel neural network for long-sequence ECG signal detection. By inputting a 10s ECG signal, the network can output the type of heart rhythm to which the signal belongs. The network achieved a classification accuracy of 95.72% in the MIT-BIH Arrhythmia database, which is higher than existing neural networks. And the size is only 314.75KB. On the other hand, this paper also proposes a layer-wise quantization method based on the Greedy Algorithm and 2 quantization modes for reducing energy and memory. It reduces the memory size of 6.83× and 15.50× in two modes respectively with only 0.33% and 2.63% classification accuracy degradation. In the future, we will consider implementing this work in the ASIC chip for heart rhythm detection.

REFERENCES

1. <https://www.texasheart.org/heart-health/heart-information-center/topics/arrhythmia/>
2. Organization W H . WHO | Noncommunicable diseases country profiles 2018[J]. Scand J Soc Med, 2018, 14(1):7-14.
3. <https://www.heart.org/en/health-topics/arrhythmia/symptoms-diagnosis--monitoring-of-arrhythmia>
4. Al-Khatib SM et al 2018 2017 aha/acc/hrs guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: a report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines and the heart rhythm society J. Am. Coll. Cardiol. 72 e91–220
5. S. Hong, Y. Zhou, M. Wu, J. Shang, Q. Wang, H. Li, and J. Xie, "Combining deep neural networks and engineered features for cardiac arrhythmia detection from ECG recordings," *Physiological Measurement,* vol. 40, p. 054009, 2019-06-04 2019.
6. P. de Chazal, R. R. B. and O. M., "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering,* vol. 51, pp. 1196-1206, 2004.
7. K. S. Park, B. H. Cho, D. H. Lee, S. H. Song, J. S. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim, "Hierarchical support vector machine based heartbeat classification using higher order statistics and hermite basis function," in *2008 Computers in Cardiology,2008 Computers in Cardiology*, 2008, pp. 229-232.
8. M. Llamedo and J. P. Martinez, "Heartbeat Classification Using Feature Selection Driven by Database Generalization Criteria," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 616-625, 2011.
9. P. Pławiak, "Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system," *Expert Systems with Applications,* vol. 92, pp. 334-349, 2018.
10. S. Parvaneh, J. Rubin, S. Babaeizadeh, and M. Xu-Wilson, "Cardiac arrhythmia detection using deep learning: A review," *Journal of Electrocardiology,* vol. 57, pp. S70-S74, 2019.
11. S. Kiranyaz, T. Ince and M. Gabbouj, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks," *IEEE Transactions on Biomedical Engineering,* vol. 63, pp. 664-675, 2016.
12. S. S. Xu, M. Mak and C. Cheung, "Towards End-to-End ECG Classification With Raw Signal Extraction and Deep Neural Networks," *IEEE Journal of Biomedical and Health Informatics,* vol. 23, pp. 1574-1584, 2019.
13. J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *IEEE Transactions on Biomedical Engineering,* vol. BME-32; BME-32, pp. 230-236, 1985-01-01; 1985-01-01 1985.
14. N. Wang, J. Zhou, G. Dai, J. Huang, and Y. Xie, "Energy-Efficient Intelligent ECG Monitoring for Wearable Devices," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 13, pp. 1112-1121, 2019.
15. Ö. Yıldırım, P. Pławiak, R. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Computers in Biology and Medicine,* vol. 102, pp. 411-420, 2018.
16. A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine,* vol. 25, pp. 65-69, 2019.
17. F. Sufi, Q. Fang, I. Khalil, and S. S. Mahmoud, "Novel methods of faster cardiovascular diagnosis in wireless telecardiology," *IEEE Journal on Selected Areas in Communications,* vol. 27, pp. 537-552, 2009.
18. L. V. Batista, E. U. K. Melcher and L. C. Carvalho, "Compression of ECG signals by optimized quantization of discrete cosine transform coefficients," *Medical Engineering & Physics,* vol. 23, pp. 127-134, 2001-01-01 2001.[1]
19. C. J. Deepu and Y. Lian, "A low complexity lossless compression scheme for wearable ECG sensors," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 449-453.
20. V. Vanhoucke, A. Senior, and M. Z. Mao. Improving the speed of neural networks on CPUs. In the Deep Learning and Unsupervised Feature Learning Workshop at NIPS, 2011.
21. M. Courbariaux, Y. Bengio and J. David, "BinaryConnect: Training Deep Neural Networks with binary weights during propagations," 2015-01-01 2015.
22. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients
23. Y. Zhou, S. Moosavi-Dezfooli, N. Cheung, and P. Frossard, "Adaptive Quantization for Deep Neural Network," 2017-01-01 2017.
24. (Y, L, Y, & P, 1998)
25. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.
26. Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yu- taka Leon Suematsu, Quoc Le, and Alex Kurakin. Large-scale evolution of image classifiers. arXiv preprint:1703.01041, 2017.
27. M. Pallavi and H. M. Chandrashekar, "Study and analysis of ECG compression algorithms,", 2016, pp. 2028-2032.
28. Minimum energy quantized neural networks
29. E. Park, S. Yoo and P. Vajda, "Value-aware Quantization for Training and Inference of Neural Networks," 2018-01-01 2018.
30. E. Park, J. Ahn and S. Yoo, "Weighted-Entropy-Based Quantization for Deep Neural Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7197-7205.
31. Pławiak, Paweł. (2017). ECG signals (1000 fragments). 10.13140/RG.2.2.21065.03685/1.
32. G. B. Moody and R. G. Mark, “The impact of the MIT-BIH arrhythmia database,” *IEEE Eng. Med. Biol., vol. 20, no. 3,* pp. 45–50, May 2001.
33. P. Pławiak, "Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals," *Swarm and Evolutionary Computation,* vol. 39, pp. 192-208, 2018.