

Received April 30, 2022, accepted May 30, 2022, date of publication June 8, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3180829

Efficient CNN Architecture on FPGA Using High Level Module for Healthcare Devices

AHMED K. JAMEIL¹ AND HAMED AL-RAWESHIDY¹, (Senior Member, IEEE)

Department of Electronic and Electrical Engineering, College of Engineering, Design, and Physical Sciences, Brunel University London, Uxbridge UB8 3PH, U.K.

Corresponding author: Ahmed K. Jameil (2006957@brunel.ac.uk)

This work was supported by Brunel University London.

ABSTRACT Modern wearable healthcare devices require new technologies with resource efficiency in terms of high performance, low energy consumption and diagnostic accuracy. In the field of artificial intelligence, the convolutional neural network (CNN) has performed as an effective algorithm. Field-programmable gate arrays (FPGAs) have been extensively utilised to construct hardware accelerators for CNNs. This paper suggests using an accelerator to create a specific 1-D CNN to classify the electrogram (ExG). ExGs used here include electrocardiogram, electroencephalogram and electromyography. The pipelined structure is designed with a register in the middle to facilitate easy data transfer. A 1-D CNN using an accelerator to categorise ExG signals implemented on Xilinx Zynq xc7z045 platform outperforms FPGA peer applications on the same platform by $1.14\times$ in terms of speed. In addition, the 1-D CNN proposed accelerator operates very efficiently due to the use of a tristate buffer in the multiplexer and the substitution of the shift for the multiplier, resulting in a resource-efficient accelerator with 161 GOP/s/W energy efficiency and 28 GOP/s/KLUT, an improvement of 1.67 over the previous model. Finally, the performance of the accelerator applied to a Xilinx Zynq xc7z045 FPGA operating at 442.948MHz was calculated, achieving 1.145 TFLOP/s.

INDEX TERMS Electrogram classification, convolutional neural network (CNN), FPGA, healthcare monitoring, hardware accelerator.

I. INTRODUCTION

Nowadays, thinking about how to use wearable Internet of Things (IoT) sensors to look at people's behaviours and determine how healthy they are is important. Wearable sensors for tracking are used in the medical field, and IoT assists in data collection via decision-making tools [1], [2]. Illnesses are often diagnosed using a cloud computing platform. Also, the huge amount of data kept and shared by many health research institutions around the world makes it hard for humans to find important information in medical data. As a result, the existing medical system continues to need a remarkable amount of time and effort from the overwhelming majority of people to provide a good medical diagnosis. In addition, the development of a wearable healthcare device with the potential of high-precision medical diagnostics is urgently required to address this issue [3]–[6].

Artificial intelligence (AI) models, surgical gadgets and mixed-reality applications may be used to diagnose and treat

illnesses more effectively than ever before. Consequently, the clinical decision support system achieves particular goals such as the detection of electroencephalograms (EEG) and electromyography (EMG). AI diagnosis is also more accurate than manual diagnosis. Moreover, machine learning (ML)-based models outperform human pathologists and imaging specialists in terms of precision. With smart diagnosis, a patient's current health status and ailment severity may be precisely identified, so that a tailored treatment plan can be established [7], [8].

Electrocardiogram (ECG) is used to visualise the heart's electrical activity. ECG contains a wealth of information due to the simplicity with which it may be monitored. It may offer vital information about the progression of a myocardial infarction, the presence of various cardiac arrhythmias and the effect of hypertension [9]. Moreover, in standard EEG, electrical activity in the cortex is captured through scalp electrodes and shown as a waveform. Implanting electrodes directly on the exposed surface of the brain is necessary to capture electrical activity from the cerebral cortex. This process includes expressions of subcortical areas in corti-

cal regions. In addition, EEG may be used to monitor the brain's functional integrity because it reflects the functional condition of the brain and as a result, aids doctors in identifying a wide range of neurological diseases. It is also critical to establish a link between scalp potentials and the underlying neurophysiology. It may detect pathological conditions such as ordinary headaches and dizziness, epilepsy, brain tumours and multiple sclerosis as well as sleep problems and movement irregularities [10]–[12].

EMG is a method for recording and analysing signals produced by the electrical activity of skeletal muscles that has been around for some time. These signals are also called myoelectric signals. EMG is utilised as an assessment tool in a variety of domains such as applied research, physiotherapy, sports training and other similar fields of study. Surface EMG provides information on the general function and conduction of muscles. To obtain myoelectric signals, electrodes are placed on the surface of the skin. When EMG is conducted on particular muscles, such as shoulder and upper trunk, or when it is recorded in monopolar mode, it is often polluted by ECG. This cardiac artefact in imaging cannot be prevented. Consequently, to extract suitable, qualitative information from an EMG signal, removing cardiac artefacts from the signal and deriving a pure EMG signal from it are necessary [11], [13], [14].

Many algorithms were previously built on morphological features and traditional signal processing techniques [15]–[24]. Fixed features used in such algorithms cannot properly discriminate between various forms of ExG because the ExG waveform and its morphological qualities fluctuate greatly depending on the situation and the patient. In addition, DL methods were recently developed to extract features automatically and improve ExG classification accuracy. DL approaches have been shown to be very adaptable and precise in the classification of ExG amongst other applications [25]. For instance, Cimtay *et al.* [26] demonstrated emotion recognition using facial expressions and EEG, and achieved 91.5% maximum accuracy and 53.8% mean. Oh *et al.* [27] proposed a hybrid model that consists of convolutional neural networks (CNN) and long-term memory (LSTM) to increase the accuracy of detection of arrhythmias. This model may detect a common arrhythmia. With variable length data, it achieved good classification accuracy (98.10%), sensitivity (97.5%) and specificity (97.5%). Alam *et al.* [28] proposed a new concept design with regard to the detection model portion. Human psychophysiological data were acquired using EMG, electrodermal activity and ECG sensors, and processed using a CNN to detect the hidden emotional state.

Preprocessing, feature extraction and classification methods are often used in conjunction to achieve ExG classification. Although LSTM is excellent at processing time domain data such as EMG, it is only capable of basic feature extraction in the context of EMG classification. When using LSTM, sophisticated preprocessing algorithms and classification are required to achieve high classification accuracy,

making decreasing resource usage throughout the implementation on the hardware side challenging [29]–[31]. In contrast to LSTM, a sophisticated preprocessing technique for the network to use it need not be developed because of the remarkable self-learning capabilities and flexibility of CNN. Thus, building a CNN-based classification model is more efficient in terms of hardware use.

In this paper, a 1-D CNN structure for the detection and classification of ExG signals (ECG beats, EMG and EEG signals) is presented. The 1-D CNN is implemented using an efficient hardware design. The developed 1-D CNN model and hardware architecture may also be used for other time series applications, such as blood pressure and diabetes monitoring. The following are the most substantial contributions of this work:

- 1) This paper is the first to design a hardware architecture using three biomedical signals from ExG in field-programmable gate array (FPGA) platform for facilitating CNN acceleration. Additionally, the proposed architecture can compute convolution for any size of input and modify the stride value.
- 2) This work discusses a 1-D CNN structure tailored to an embedded application. By using Python, it produces an ensemble of output from 1-D CNN layers for each ExG with 99% accuracy, and the 1-D CNN recognises the ExG signal.
- 3) This design maximises the use of hardware resources whilst minimising the accuracy loss of ExG categorisation.
- 4) This work designs a pipelined processing unit array to achieve great performance and efficiency. It also includes a sign bit in each processor unit, which not only minimises power consumption but also lowers the cost of hardware resources. As a result, the proposed design achieved a high performance of 1.145 TFLOP/s at 442.948 MHz and 1.068 KLUT resource utilisation.
- 5) The design accurately identifies the ExG signal using FPGA Xilinx and attains a higher speed than classification of just one type.

The remainder of this paper is arranged as follows: Section II discusses the features of ECG, EEG and EMG signals as well as the history of CNN algorithm and its applications. Section III presents the proposed accelerator of 1-D CNN and explains in detail how to design the new architecture. Section IV focuses on the proposed structure of the 1-D CNN and other algorithms. Section V discusses the results and compares them with those of other models. Section VI summarises the conclusion.

II. RELATED WORKS

CNNs have seen much success in deep learning in recent years [7]. Using CNNs as the basis for the vector convolution calculation approach, new concepts for learning with high classification accuracy have been discovered. This section includes several relevant research works, such as EEG, EMG

and ECG signals as well as classification learning research methods based on CNN and other algorithms, which can differentiate between CNN and other algorithms in terms of architecture, accuracy, speed of diagnosis and classification in healthcare monitoring.

A. EEG SIGNAL

Social integration can be substantially improved by using EEG-based emotion ratings for patients with early-stage Alzheimer's disease and neurological disorders. Moreover, emotions have traditionally been classified via the use of software running on computers and working without an internet connection. However, these classifiers can be worn, which is important to enhance the social life of patients. This level of wearability must be achieved by the deployment of low-power hardware resources that enable near real-time classification and long durations of operation. Gonzalez *et al.* [31] proposed a hardware CNN called BioCNN that is employed to optimise EEG-based emotion detection and other biomedical applications. They used Digilent Atlys Board in conjunction with a low-cost Spartan-6 FPGA to accomplish the training technique. Their results [31] revealed 100 MHz speed, 26.229 KLUT resource consumption and 0.0629% resource consumption efficiency.

Two key study issues in the field of EEG categorisation are the detection of epileptic seizures and the identification of emotional states. Through the development of a real-time, energy-efficient processor to conduct on-device seizure detection, informing others in the immediate vicinity or immediately preventing the seizure by providing instant stimulation will be feasible. Accurate seizure detection requires precision to be safe [25]. Sahani *et al.* [32] illustrated that scalp multichannel signalling and electroencephalography are both effective methods to detect epileptic seizures in real time. Additionally, they developed a novel architecture to extract additional features with great accuracy and speed, and implemented it using FPGA platform Virtex-5. The architecture [32] showed that it can detect and identify epileptic episodes in a steady, reliable manner. Speed was 86.73 MHz, and resource usage was 11.963 KLUT.

B. EMG SIGNAL

In recent years, EMG processors have received a great deal of interest because they are often employed in gesture recognition applications. To maximise classification accuracy whilst minimising power consumption, wearable devices are generally used for gesture recognition. In [33], a low-power embedded EMG acquisition and gesture recognition system was proposed. Software and hardware multilevel design optimisation was emphasised. In addition to EMG sensors, inertial and pressure sensors have been utilised to increase gesture identification and motion tracking accuracy.

The development of specialised CNN accelerators has opened up new possibilities for edge healthcare and biomedical applications [34]. Franco *et al.* [35] proposed a set of readings to analyse surface-EMG (sEMG) to study how the nervous system modulates muscle activity. The

researchers built an FPGA-based real-time Non-Negative Matrix Factorization (NMF) processor that extracts muscle synergies from 8-electrode EMG recordings and feeds them to an SVM classifier. It was implemented on FPGA platform. In addition, their results [35] revealed 87.74 MHz speed, 38.836 KLUT resource consumption and 7.2 W power consumption. Mostafa *et al.* [36] proposed an architecture design and execution for estimating the desired clench strength of the hand using EMG signal and implemented this using Xilinx's XC7Z020 platform. They also showed that the proposed architecture can be used for any application related to prosthetics. In accordance with the findings of [36], speed was 388.20 MHz, resource usage was 4.379 KLUT and power consumption was 0.344 W.

C. ECG SIGNAL

In medicine, an ECG device is used to record the electrical activity of the heart's pulse to diagnose various forms of cardiac disease. Electrodes are placed on the patient's body to attach the ECG [37], [38]. The development of wearable sensors for healthcare monitoring has the ability to read and analyse different parts of the ECG [39]. Also, real-time monitoring systems use a lot of different methods to identify ECGs [40]. In recent years, IoT has been used to monitor patients and their health remotely. By training on a data set such as cardiology, AI algorithms are also used to classify and identify diseases with high speed and accuracy. Moreover, appropriate processors are designed for these data to expedite disease diagnosis [41]–[43]. Lu *et al.* [44], suggested a hardware design for an integrated ECG classification using a 1-D CNN with global average pooling. They built the efficient hardware design on FPGA Xilinx Zynq and delivered an average rate of 25.7 GOP/s at 200 MHz with 1538 LUT resource usage and optimised resource efficiency by more than thrice that of the non optimised scenario. The completely pipelined processing unit array meant to boost calculation speed. The accuracy of ECG beat classification was 99.10%.

The design in a study by Guo *et al.* [45] used the specifications of Angel-Eye, a programmable, adaptable CNN accelerator architecture that includes a data quantisation technique and a compilation tool. The use of a data quantisation approach may help lower bit width from 16 bits to 8 bits whilst maintaining minimal accuracy loss. The compilation tool efficiently adapts a certain CNN model to the available hardware. According to testing on Zynq XC7Z045 platform, compared with its equivalent FPGA running on the same platform, Angel-Eye operated at a faster hardware speed of 150 MHz with resource utilisation of 183 KLUT and power usage of 9.63 W. In [45], the model provided comparable performance with up to 16% more energy economy. Gong *et al.* [46] developed a novel FPGA-based accelerator architecture that operated synchronously as a pipeline of instructions. The model also generated focused optimisation, which may be used to reach the highest possible level of computing efficiency. The CNN model was used to test the performance of this accelerator on a variety of platforms,

including Xilinx Zynq-7020 and Virtex FPGA. The model obtained 200 MHz speed with 2.15 W power consumption and 38.136 KLUT.

D. CNN MODEL AND FPGA

Pattern classification and data mining studies have been enhanced by neural network (NN) success. Many ML tasks that previously depended largely on handmade feature engineering have lately been transformed by end-to-end deep learning models such as CNN [47]. A 1-D CONV layer is generated from numerous computational layers organised as directed acyclic graphs. Each layer extracts a feature map, which is an abstraction of the data supplied by the preceding layer. The output of result y_n is shown as

$$y_n = b_n + \sum_{k=0}^{k-1} w_{nk}x_k \quad (1)$$

where w_{nk} and x_k are the kernel weights and the feature map data, respectively, b_n is the bias and k is the number of feature maps provided. Using the output of result y_n as a starting point, output Y_k can be represented as

$$Y_k = f(y_n) \quad (2)$$

Additionally, f represents the activation function, which is commonly rectified linear unit (ReLU) in CNN.

$$\text{ReLU}(x) = \max(0, x) \quad (3)$$

Convolution, pooling and fully connected layers are the most popular layers. 1-D CNNs have been frequently utilised in medical and healthcare applications. A 1-D CNN employs convolutional layers, which use spatial filters to promote spatial invariance. When convolutional layers are included in 1-D CNNs, spatial filters are used to enhance spatial invariance, which is beneficial. Pool layers down-sample input feature maps spatially by dividing them into subregions and merging their values into a single value. The pool operators' max-pooling and average-pooling employ maximum and average values for each subregion. The fully connected (FC) layers link all the neurons in the previous layer to every single neuron in the following layer, forming a network of connections. To build a relational representation of these characteristics for each class in the classifier detection set, FC layers join the features retrieved by the CONV layers and combine them into a single entity. Finally, a SoftMax classifier is applied to the outputs of the previous FC layer to obtain normalised class probabilities for different classes in the final layer.

In addition, FPGA acceleration for CNNs has received much interest. Using an appropriately designed FPGA accelerator for CNN, the full capability of the parallelism of low latency and fast speed can be achieved because of the high-performance, high-speed and low-power consumption needs of various applications. FPGAs are widely used as cost-effective options in many industries. Furthermore, the recon durability of FPGAs enables them to adapt quickly to new CNN designs [48]. Compared with CPU or GPU, FPGA

has better energy efficiency. Making a high-performance FPGA accelerator is a lengthy process that typically entails many steps, including parallel architectural discovery, memory bandwidth optimisation, area and timing tuning, and software–hardware interface creation. Consequently, automatic compilers were developed for FPGA CNN accelerators, in which the hardware description of target accelerators can be generated automatically based on parametric templates, and design space exploration can be simplified into parameter optimisation with respect to network structure and hardware resource constraints [49]. In this work, an FPGA accelerator is designed for CNN using three ExGs. The FPGA accelerator is designed with fewer hardware and lesser complexity, which increases classification speed and accuracy. Consequently, power consumption decreases.

III. THE PROPOSED ACCELERATOR OF 1-D CNN

Signal flow graph is a technique for discrete time modelling in systems that illustrates the iterative process of data processing or classification. This approach is used to explain the 1-D CNN CONV layer and how the processing element (PE) is designed in relation to the iterations of data processing and classification as depicted in Figure 1. The solution achieved by the multiplication and addition operations of the 1-D CNN CONV layer is shown in Figure 1. This approach is represented by variable w_{nk} and h_{ki} , which are the concatenation of the kernel weights and the feature map of the input data, respectively, with a bias b_n . To reduce the amount of hardware required for 1-D CNN, the multiplication is shifted to the left. Continuous collection operations follow, which necessitates the use of a register (R). After completing the iterative data processing cycle in R, it is then collected with b_n . Finally, the output of y_k is the final result.

In addition, the developed architecture of 1-D CNN reduces the complexity of multiplication by using shift operations. The shift in Figure 2 is represented by the symbol im , which can be stated as shown in the equations below based on Eq. (1).

$$X_k = h_{k,0} + h_{k,1} \times 2^m + h_{k,2} \times 2^{2m} + \dots + h_{k,i} \times 2^{im} \quad (4)$$

Then

$$X_k = h_{k,i} \times 2^{im} \quad (5)$$

where i^{th} is the partition of kernel weights of the 1-D CNN with the length of shift $m = k/i$. In the next step, Eq. (5) is substituted into Eq. (1) and yields.

$$y_n = b_n + \sum_{k=0}^{k-1} \sum_{i=0}^{i-1} h_{k,i} \times 2^{im} \times w_{nk} \quad (6)$$

The ultimate shape of the proposed architecture is determined by Eq. (6). The PE in Figure 2 consists of an XOR gate that serves as a selector to examine the sign bit. It employs a tristate buffer instead of other gates in the multiplexer to decrease the number of devices needed.

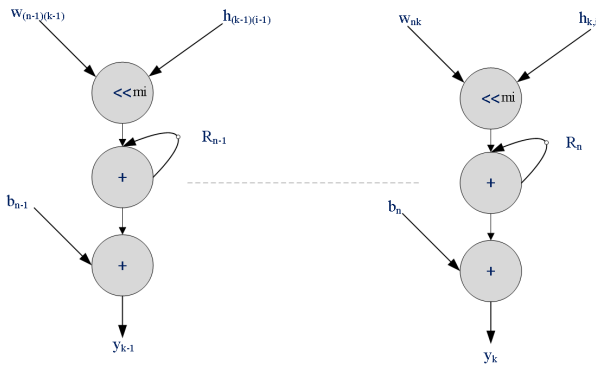


FIGURE 1. Signal flow graph of a 1-D CNN CONV layer.

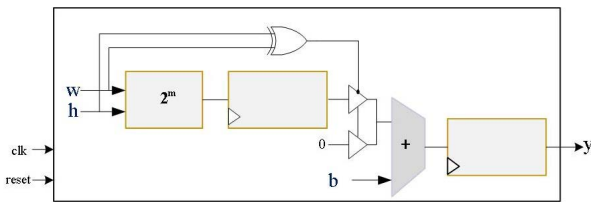


FIGURE 2. Structure of a PE.

Figure 3 depicts the architecture of the proposed CNN accelerator, in which data are stored in off-chip memory prior to the start of classification. Kernels and weights are extraordinarily huge in terms of data size and are consequently stored in our implementation’s off-chip memory. However, on-chip buffer is often inadequate for caching all of the parameters and data for state-of-the-art 1-D CNNs. As a consequence, off-chip memory is utilised to store all of the network settings as well as the results of each layer. The on-chip buffer approach was chosen to effectively feed data to the PE arrays based on two factors. First, by preloading cores and weights from off-chip memory to on-chip buffer using the data bus, PE arrays can access the required data at high speeds. Second, this approach loads a collection of data from off-chip memory rather than a single datum at a time, which maximises memory bandwidth utilisation. The PE array is connected to the off-chip memory via the on-chip buffer using a data bus. This enables parallelization of data input/output and computation. Additionally, the output buffer provides interim results to the PE array and max pooling if an output channel requires more than one cycle of computation. This work introduces a 1-D CNN description interface for data management. In addition, this allows the user to make full use of the on-chip buffer.

In this study, the 1-D CNN is initially trained using 32-bit floating-point data to determine the data and parameter ranges of each layer. The bit width used in the proposed architecture is 16 bits, and the stride is 1. The data are divided into three categories in the buffer prior to inserting the data into the PE array. The PE array uses a pipeline structure by placing a shift register between the layers of the PE array. Moreover, the shift registers function as a form of local memory for previously obtained values. The pipeline structure is used in the proposed architecture to achieve high classification

efficiency and increase speed. After storage in the PE array, the data are transmitted to the pipelined adder tree, where they are utilised to complete the calculation. The adder tree is selected because it produces a high-quality output with a low critical path latency.

The design of max pooling and its connection to the output buffer is shown in Figure 3. A MUX is placed in front of the max pooling to select the verified input data to the associated max pool to support varying convolution strides and pool size scenarios. In the suggested design of the pooling, the tristate buffer mux structure is utilised instead of other gates to decrease the number of hardware used, and a selector (enable) is employed to control the output data centrally. A comparator and a useful feedback register are used in the max pool to save the final comparator output. According to the suggested architecture of 1-D CNN, the results for all layers are selected by max pooling and then sorted.

IV. PROPOSED STRUCTURE OF 1-D CNN AND OTHER ALGORITHMS

A. WORKFLOW OF THE SYSTEM

Figure 4 depicts the workflow followed throughout the system building phase. The collected data set is first saved in a database for simple retrieval and analysis. Then, preprocessing procedures such as padding, reshaping and resampling are performed on the stored data. Next, the data are divided into two sections, 1) testing data and 2) training data, which are employed in the model-building step. The model creation phase has two parts, 1) model evaluation and 2) model training. The training data are used to train the model, and the testing data are used to evaluate the model’s performance during the evaluation phase. Then, the ensemble is applied to the three models to unify the findings and output, and identify which mode is the best. To select the best model, this step is performed 10 times using various algorithms of ML. The system is now ready to accept new data samples for classification after storing the optimum model. Finally, the algorithm is implemented on FPGAs and used to create an accelerator for classification.

B. UTILISED DATA

In this paper, the ExG signal, which contains ECG, EEG and EMG signals, was utilised because the signal characteristics of these three types are very similar as shown in Figure 5. The ECG signal data set from the UCI Machine Learning Repository was used [50]. There are various variables in this data set, as well as a target state of heart disease or not having heart disease. The ECG has a sample size of 303 patients, 4242 parameters, and 76 features, but in published research, only 14 of these are used. The sampling frequency of the ECG signal was set at 100 Hz in this work. The data set of ECG signal contains: age in years; sex (1 = male, 0 = female); type of chest pain; resting blood pressure; serum cholesterol; fasting blood sugar (1 = true, 0 = false); resting electrocardiographic results; maximum heart rate achieved;

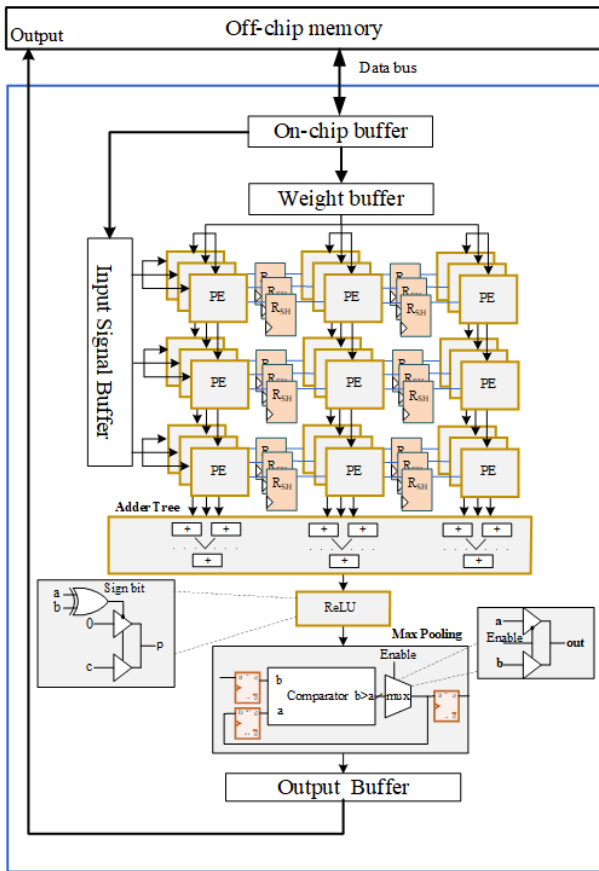


FIGURE 3. Architecture of proposed CNN accelerator.

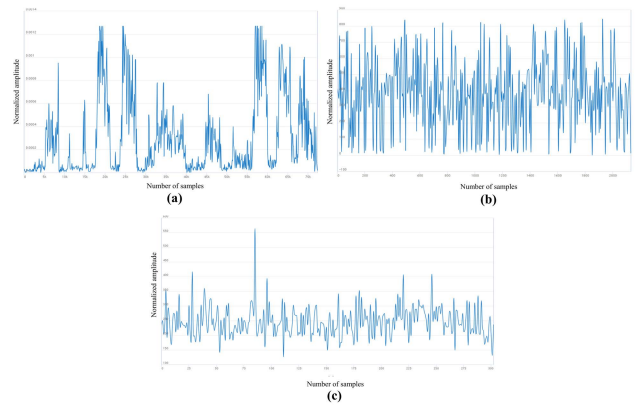


FIGURE 5. ExG signals for a) EMG, b) EEG and c) ECG.

as melancholy (negative), joy (positive) and neutral (regular) by the symbols 1, 2, and 3, respectively. The EEG signal has a sample size of 1492 number of instances, 2548 features, and the number of parameters is 3801616 with a sampling frequency of 150Hz. Finally, movie excerpts from various films were used to illustrate each of the three instances of emotion. The UCI Machine Learning Repository’s EMG signal data set was used [53]. The EMG data set utilised in this paper contained a data set used to monitor human activity on volunteers in states of normal and aggressive actions on the body as well as the state of effect on them, utilising eight channels attached to their bodies. The EMG signal has a sample size of 10,000 number of instances, 8 features, and the number of parameters used is 723220. In this case, the sample frequency of the EMG signal was set at 200 Hz. For simplicity of classification, this sheet of EMG signal data was encoded from 0 to 6, with channel number 7 for the data carrier sensor.

V. RESULTS AND DISCUSSION

The proposed model was tested with three signal data sets (ECG, EEG and EMG). As mentioned above, the proposed model was tested with four different algorithms to validate it and select the best algorithm performance through the highest possible classification accuracy of the three signal data sets of ExG. The four techniques utilised were stochastic gradient descent (SGD), naive Bayes (NB), support vector machine (SVM) and CNN. In this section, the performance of the proposed model for each algorithm is also evaluated, and the best algorithm for implementation on hardware FPGA is identified. The FPGA implementation model including a 1-D CNN algorithm that is the most effective in this model for ExG signal processing is discussed in the second part.

A. ANALYSIS OF TRAINING AND EVALUATION

In this section, the main parameters of each algorithm used, the structure of the model and the metrics used to evaluate the performance of the proposed model are discussed. Also, ECG data sets are split into training, validation, and testing. With regards to the ECG, 70% of the data set was chosen at random and positioned in the training set for the purpose of

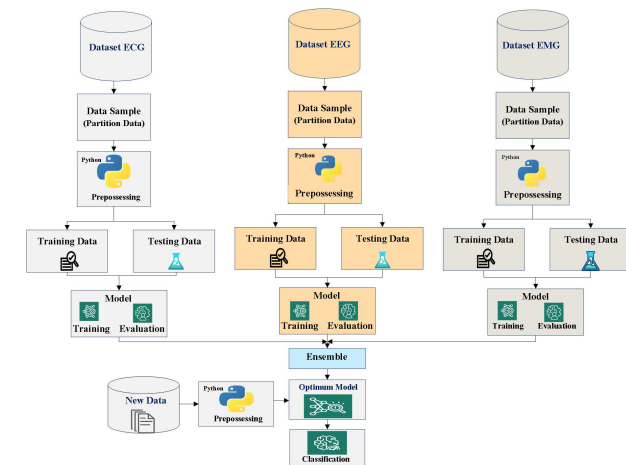


FIGURE 4. Proposed structure and system workflow for ExG signal classification.

Exercise-induced angina (1 = yes, 0 = no); ST depression induced by exercise relative to rest; Normal, a fixed defect, and a reversible defect are denoted by 3, 6, and 7, respectively; target, whether you have a sickness (1 = yes, 0 = no).

The data set for EEG signals in [51], [52] used the brain-wave data set for processing. It also utilised dry electrodes to represent each positive, negative or neutral state that participants encountered. In this paper, sentiments were classified

obtaining well-trained parameters. To assess the 1-D CNN's performance, the remaining data is split evenly by 15% for the validation and test sets. Additionally, by separating the EMG and EEG data sets using the same method that was used to compute the split ratio of the ECG data set. Accuracy measures how often a model successfully classifies data samples, as shown in the equation below for the evaluation of the module of 1-D CNN and other algorithms based on true negative (TN), true positive (TP), false positive (FP), and false negative (FN).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

The accuracy attained by each algorithm based on the NN of the proposed model is shown in Figure 6, and the 1-D CNN acquired 99% maximum level of accuracy. The accuracies of the four models were also compared, as shown in Figure 7. The graph clearly shows that 1-D CNN is more accurate than the other models. This method was used with 100 iterations with 20 trains.

The main classification metrics were calculated after collecting the confusion matrix values using the classification report consisting of recall, F1-score and precision. Figure 8 describes the performance of the proposed model of each algorithm. Precision is the accuracy of positive predictions, as demonstrated in Eq. (8) and Figure 8(c).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

Recall is the proportion of positive samples accurately recognised from the real positives, as shown in Eq.(9), is referred to as recall. The models used have lower scores than the 1-D CNN model, but all scores are excellent, as shown in Figure 8(d).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

F1-score is the harmonic mean of precision and recall, as stated in Eq. (10). Each model was assessed with regard to F1, as demonstrated in Figure 8 (b), indicating that the proposed models are excellent, but 1-D CNN is the best.

$$F1 = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \quad (10)$$

The receiver operating characteristic curve and the area under the curve (AUC) are two probability models used to compare the TP rate to the FP rate at various thresholds through a graphical representation sensitivity against specificity, as shown in Figure 8(a). AUC indicates the classifier's ability to discriminate between distinct classes. The receiver operating characteristic curve contains three main evaluation values: A value close to 1 indicates that the classifier is performing well, a value close to zero indicates that the model is 100% incorrect and classifying in reverse, and a value close to 0.5 indicates that the classifier is only guessing. The relationship between FP rate (specificity) and TP rate (sensitivity) is depicted in Figure 9. All curves close to 1 indicate that

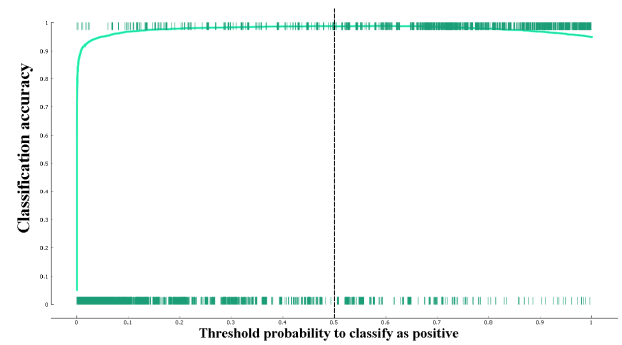


FIGURE 6. Accuracy of a 1-D CNN with threshold probability to classify as positive.

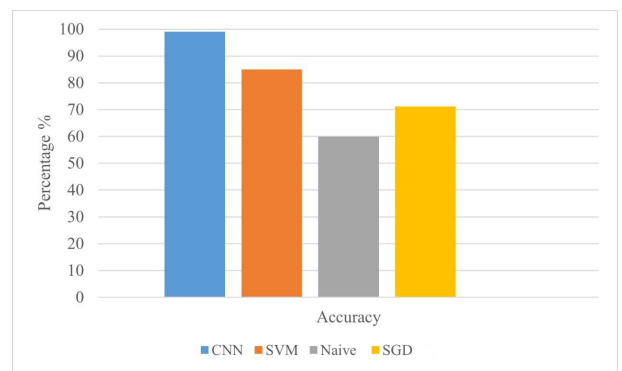


FIGURE 7. Model accuracy of CNN, SVM, NB and SGD.

the proposed model classifier performs well, except for the SGD model, which is only guessing because its curve is close to 0.5.

B. ANALYSIS OF IMPLEMENTATION ON HARDWARE

The proposed architecture of 1-D CNN was designed on Xilinx Zynq Xc7z045 platform. The same assessment carried out on the testing set on PC platform was similarly performed on hardware. In terms of performance and characteristics, Table 1 highlights the comparison between the proposed 1-D CNN accelerator based on FPGA platform and other accelerators based on the kind of data processed on the accelerator to meet the standards and the specifications. A 2-D CNN and other models were used instead because no 1-D CNN accelerators were used during the comparison. Studies in the area of EMG and EEG in 1-D CNN accelerator are limited, and the studies were selected, as stated in Table 1. The proposed accelerator achieved a high speed of 442.948 MHz compared with the highest speed of 388.29 [36] prior to this work, as shown in Figure 10(a).

Lu *et al.* [44] proposed a layout of the accelerator based on CNN that achieves the highest efficiency within the use of hardware resources of 16.71 GOP/s/kLUT. However, the proposed architecture of this work improved resource efficiency to 28 GOP/s/kLUT and is now the highest one, surpassing the previous 16.71 GOP/s/kLUT. The results demonstrate that

TABLE 1. Comparison between CNN accelerator on Zynq xc7z045 performance with other fpga accelerators.

Ref. No.	[31]	[32]	[35]	[36]	[44]	[45]	[46]	Our Work
Method	2-D CNN	CNN	SVM	ANN	1-D CNN	VGG-16	LeNet-5	1-D CNN
FD ^a	Spartan-6	Virtex-5	Zy ^J xc7z020	Zy xc7z020	Zy xc7z045	Zy XC7Z045	Zy XC7Z020	Zy xc7z045
Power(W)	1.59	1.59	7.2	0.344	-	9.63	2.15	0.176
MHz ^b	100	86.73	87.74	388.20	200	150	200	442.948
GOP/s	1.65	-	-	-	25.6	187.8	76.48	30
Data size(bit)	Fixed 16	Fixed 16	Fixed 8	Fixed 16	Fixed 16,32 Fp ^e	Fixed 16,32 Fp	Fixed 16	Fixed 16,32 Fp
Parameter	143662	4097	110000	-	11065	50.15M	60840	3878178
BRAM	10	-	91	25	12	486	242	50
KLUT	26,229	11,963	38,836	4,379	1,538	183	38,136	1,067
DSP	32	-	148	25	80	780	205	9
CT ^d	EEG	EEG	EMG	EMG	ECG	ECG	ECG	ExG ^c
GOP/S/KLUT	0.0629	-	-	-	16.71	0.749	2.005	28
TFLOP/s	-	-	-	-	358,907 ^g	2,262 ^g	-	1.145

^a FamilyDevice; ^b Frequency; ^c (ECG, EEG, EMG); ^d ClassificationTask; ^e FloatingPoint; ^f Zynq; ^g GFLOP/s

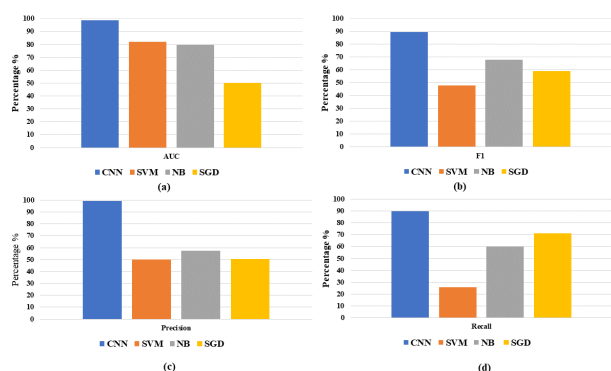


FIGURE 8. Comparison of four models in terms of (a) Area under the curve, (b) F1-score, (c) Precision and (d) Recall.

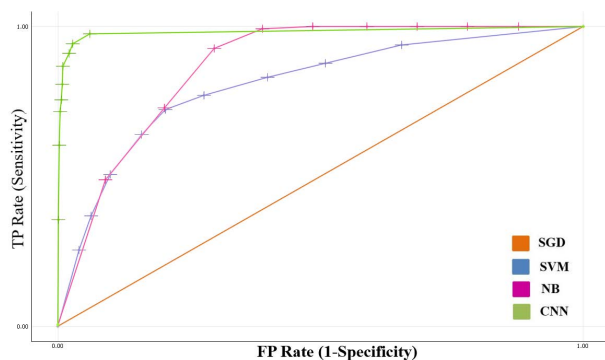


FIGURE 9. The receiver operating characteristic curve of four models.

our solution can reach a peak performance of 161 GOP/s/W energy efficiency on Zynq xc7z045 processor, which is much better than the previously reported results for other techniques. Several ExG biosignals, such as EMG, EEG and ECG, were described in this paper. The approach was implemented on FPGA platform and enabled multiclass learning and classification of the signals. Several methods were also employed to decrease the number of devices used. For example, in the multiplexer, a tristate buffer was used to reduce the number of resource utilisations that contain sign-enabled utilisation for central control. Moreover, in this work, the lowest

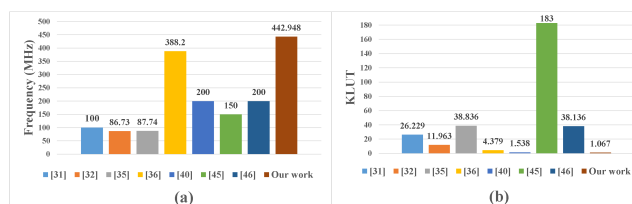


FIGURE 10. Description of proposed 1-D CNN accelerator: a) Frequency comparison between the proposed structure and other models b) Resource utilisation comparison (KLUT).

value of resource utilisations was 1.067 KLUT, as shown in Figure 10 (b). In addition, the sign bit was considered because the XOR gate was configured as a selector to control the sign bit.

The bandwidth requirement of off-chip memory is determined directly by the access data width of the memory. So, the bandwidth requirement of the memory in off-chip memory is directly proportional to the access data width of the memory. When accessing 32-bit data, the bandwidth requirement of the memory is 1.8 GB/s, but when accessing 512-bit data, the bandwidth requirement of the memory rises to 28.4 GB/s. In this work, the access data width is set to 512-bit in order to optimise memory bandwidth. Kernels and weights have been meticulously structured to accommodate the 512-bit access mode that is operating at 442.948 MHz.

At the end of this study, the peak single-precision throughput (TFLOP/s) of mainstream FPGA devices is measured by the method proposed in [54]. The proposed design has a peak single-precision throughput of 1.145 TFLOP/s, which is higher than the earlier designs [44], [45], which had 358.907 GFLOP/s and 2.262 GFLOP/s, respectively.

VI. CONCLUSION

This paper proposes a new model for classifying aggregated ExG signals consisting of ECG, EEG and EMG. When the proposed model is applied to four algorithms to ensure that it works properly, the average accuracy of 1-D CNN algorithm can reach 99%, which is higher than that of the rest of the models. In addition, the proposed model with the highest average accuracy of a 1-D CNN accelerator is

implemented to classify ExG signals in wearable healthcare devices. The pipelined structure is designed with a register in the middle to facilitate easy data transfer. The proposed 1-D CNN accelerator works very efficiently due to using a tristate buffer in the multiplexer and replacing the multiplier by shift, which results in a resource-efficient accelerator with 161 GOP/s/W energy efficiency and 28 GOP/s/KLUT that is improved by $1.67\times$ compared with the previous model. Also, the floating-point processing parts have a peak rate of 1.145 TFLOP/s, while the off-chip memory has a capacity of 28.4 GB/s. A 1-D CNN with an accelerator for classifying ExG signals implemented on Xilinx Zynq xc7z045 platform outperforms FPGA peer applications on the same platform by $1.14\times$ in terms of speed. In the future, new technologies that improve classification performance, speed and efficiency whilst using less hardware resources will be investigated. An accelerator that enables the integration of additional applications, speed evaluation and execution at any moment with minimal power usage will be developed. The CNN accelerator will be extended to other biological applications using a variety of signalling and processing approaches.

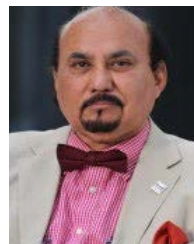
REFERENCES

- [1] E. Kristiani, C.-T. Yang, C.-Y. Huang, P.-C. Ko, and H. Fathoni, "On construction of sensors, edge, and cloud (iSEC) framework for smart system integration and applications," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 309–319, Jan. 2020.
- [2] Y. Sasaki, "A survey on IoT big data analytic systems: Current and future," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1024–1036, Jan. 2021.
- [3] G. Zhang, S. Ni, and P. Zhao, "Learning-based joint optimization of energy delay and privacy in multiple-user edge-cloud collaboration MEC systems," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1491–1502, Jan. 2021.
- [4] F. Sun, W. Zang, H. Huang, I. Farkhatdinov, and Y. Li, "Accelerometer-based key generation and distribution method for wearable IoT devices," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1636–1650, Feb. 2020.
- [5] H. Wang, Z. Qu, Q. Zhou, H. Zhang, B. Luo, W. Xu, S. Guo, and R. Li, "A comprehensive survey on training acceleration for large machine learning models in IoT," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 939–963, Jan. 2021.
- [6] W. Li, G.-G. Wang, and A. H. Gandomi, "A survey of learning-based intelligent optimization algorithms," *Arch. Comput. Methods Eng.*, vol. 28, no. 5, pp. 3781–3799, Aug. 2021.
- [7] Y. Liu, J. Wang, J. Li, S. Niu, and H. Song, "Machine learning for the detection and identification of Internet of Things devices: A survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 298–320, Jan. 2021.
- [8] R. F. Mansour, A. E. Amraoui, I. Nouaouri, V. G. Diaz, D. Gupta, and S. Kumar, "Artificial intelligence and Internet of Things enabled disease diagnosis model for smart healthcare systems," *IEEE Access*, vol. 9, pp. 45137–45146, 2021.
- [9] R. B. Reilly and T. C. Lee, "Electrograms (ECG, EEG, EMG, EOG)," *Technol. Health Care*, vol. 18, no. 6, pp. 443–458, 2010.
- [10] E. H. T. Shad, M. Molinas, and T. Ytterdal, "Impedance and noise of passive and active dry EEG electrodes: A review," *IEEE Sensors J.*, vol. 20, no. 24, pp. 14565–14577, Dec. 2020.
- [11] J. A. Mucarquer, P. Prado, M.-J. Escobar, W. El-Deredy, and M. Zanartu, "Improving EEG muscle artifact removal with an EMG array," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 815–824, Mar. 2019.
- [12] M. Li and G.-G. Wang, "A review of green shop scheduling problem," *Inf. Sci.*, vol. 589, pp. 478–496, Apr. 2022.
- [13] B. S. Darak and S. M. Hambarde, "A review of techniques for extraction of cardiac artifacts in surface EMG signals and results for simulation of ECG-EMG mixture signal," in *Proc. Int. Conf. Pervasive Comput. (ICPC)*, Jan. 2015, pp. 1–5.
- [14] S. A. Raurale, J. McAllister, and J. M. D. Rincon, "EMG biometric systems based on different wrist-hand movements," *IEEE Access*, vol. 9, pp. 12256–12266, 2021.
- [15] T. Teijeiro, P. Félix, J. Presedo, and D. Castro, "Heartbeat classification using abstract features from the abductive interpretation of the ECG," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 409–420, Mar. 2016.
- [16] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.
- [17] K. Minami, H. Nakajima, and T. Toyoshima, "Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 2, pp. 179–185, Feb. 1999.
- [18] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ECG complexes using Hermite functions and self-organizing maps," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 838–848, Jul. 2000.
- [19] L.-Y. Shyu, Y.-H. Wu, and W. Hu, "Using wavelet transform and fuzzy neural network for VPC detection from the Holter ECG," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1269–1273, Jul. 2004.
- [20] O. T. Inan, L. Giovannardi, and G. T. A. Kovacs, "Robust neural-network-based classification of premature ventricular contractions using wavelet transform and timing interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2507–2515, Dec. 2006.
- [21] P. Bera, R. Gupta, and J. Saha, "Preserving abnormal beat morphology in long-term ECG recording: An efficient hybrid compression approach," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2084–2092, May 2019.
- [22] D. A. Coast, R. M. Stern, G. G. Cano, and S. A. Brillner, "An approach to cardiac arrhythmia analysis using hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 9, pp. 826–836, Sep. 1990.
- [23] P. De Chazal and R. B. Reilly, "A patient-adapting heartbeat classifier using ecg morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2535–2543, 2006.
- [24] W. Jiang and S. G. Kong, "Block-based neural networks for personalized ecg signal classification," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1750–1761, 2007.
- [25] Y. Wei, J. Zhou, Y. Wang, Y. Liu, Q. Liu, J. Luo, C. Wang, F. Ren, and L. Huang, "A review of algorithm & hardware design for AI-based biomedical applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 145–163, Apr. 2020.
- [26] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.
- [27] S. L. Oh, E. Y. K. Ng, R. S. Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Comput. Biol. Med.*, vol. 102, no. 1, pp. 278–287, Nov. 2018.
- [28] M. G. R. Alam, S. F. Abedin, S. I. Moon, A. Talukder, and C. S. Hong, "Healthcare IoT-based affective state mining using a deep convolutional neural network," *IEEE Access*, vol. 7, pp. 75189–75202, 2019.
- [29] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based auto-encoder model for ECG arrhythmias classification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1232–1240, Apr. 2019.
- [30] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-based ECG classification for continuous monitoring on personal wearable devices," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 515–523, Feb. 2019.
- [31] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel, "BioCNN: A hardware inference engine for EEG-based emotion detection," *IEEE Access*, vol. 8, pp. 140896–140914, 2020.
- [32] M. Sahani, S. K. Rout, and P. K. Dash, "Epileptic seizure recognition using reduced deep convolutional stack autoencoder and improved kernel RVFLN from EEG signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 3, pp. 595–605, Jun. 2021.
- [33] S. Benatti, F. Casamassima, B. Milosevic, E. Farella, P. Schönlé, S. Fateh, T. Burger, Q. Huang, and L. Benini, "A versatile embedded platform for EMG acquisition and gesture recognition," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 5, pp. 620–630, Oct. 2015.
- [34] P. Schonle, F. Schulthess, S. Fateh, R. Ulrich, F. Huang, T. Burger, and Q. Huang, "A DC-connectable multi-channel biomedical data acquisition ASIC with mains frequency cancellation," in *Proc. ESSCIRC (ESSCIRC)*, Sep. 2013, pp. 149–152.
- [35] G. Franco, P. Cancian, L. Cerina, E. Besana, N. Beretta, and M. D. Santambrogio, "FPGA-based muscle synergy extraction for surface EMG gesture classification," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2017, pp. 1–4.
- [36] S. S. Mostafa, M. A. Awal, M. Ahmad, and F. Morgado-Dias, "Design of sEMG-based clench force estimator in FPGA using artificial neural networks," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15813–15823, Oct. 2020.

- [37] L. Guo, Z. Chen, D. Zhang, K. Liu, and J. Pan, "Age-of-information-constrained transmission optimization for ECG-based body sensor networks," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3851–3863, Mar. 2020.
- [38] H. B. Seidel, M. M. A. da Rosa, G. Paim, E. A. C. da Costa, S. J. M. Almeida, and S. Bampi, "Approximate pruned and truncated Haar discrete wavelet transform VLSI hardware for energy-efficient ECG signal processing," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 5, pp. 1814–1826, May 2021.
- [39] B. M. Li, A. C. Mills, T. J. Flewellin, J. L. Herzberg, A. S. Bosari, M. Lim, Y. Jia, and J. S. Jur, "Influence of armband form factors on wearable ECG monitoring performance," *IEEE Sensors J.*, vol. 21, no. 9, pp. 11046–11060, May 2021.
- [40] M. Wasimuddin, K. Elleithy, A. Abuzneid, M. Faezipour, and O. Abuzagheh, "Stages-based ecg signal analysis from traditional signal processing to machine learning approaches: A survey," *IEEE Access*, vol. 8, pp. 177782–177803, 2020.
- [41] Q. Dong, R. S. Downen, B. Li, N. Tran, and Z. Li, "A cloud-connected multi-lead electrocardiogram (ECG) sensor ring," *IEEE Sensors J.*, vol. 21, no. 14, pp. 16340–16349, Jul. 2021.
- [42] A. M. Rateb, "A fast compressed sensing decoding technique for remote ECG monitoring systems," *IEEE Access*, vol. 8, pp. 197124–197133, 2020.
- [43] J. Qian, P. Tiwari, S. P. Gochhayat, and H. M. Pandey, "A noble double-dictionary-based ECG compression technique for IoTH," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10160–10170, Feb. 2020.
- [44] J. Lu, D. Liu, Z. Liu, X. Cheng, L. Wei, C. Zhang, X. Zou, and B. Liu, "Efficient hardware architecture of convolutional neural network for ECG classification in wearable healthcare device," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 7, pp. 2976–2985, Jul. 2021.
- [45] K. Guo, L. Sui, J. Qiu, J. Yu, J. Wang, S. Yao, S. Han, Y. Wang, and H. Yang, "Angel-eye: A complete design flow for mapping CNN onto embedded FPGA," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 1, pp. 35–47, Jan. 2017.
- [46] L. Gong, C. Wang, X. Li, H. Chen, and X. Zhou, "MALOC: A fully pipelined FPGA accelerator for convolutional neural networks with all layers mapped on chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 11, pp. 2601–2612, Jul. 2018.
- [47] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2020.
- [48] A. Saidi, S. B. Othman, M. Dhoubi, and S. B. Saoud, "FPGA-based implementation of classification techniques: A survey," *Integration*, vol. 81, pp. 280–299, Nov. 2021.
- [49] T. Yuan, W. Liu, J. Han, and F. Lombardi, "High performance CNN accelerators based on hardware and algorithm co-optimization," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 1, pp. 250–263, Jan. 2020.
- [50] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Amer. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.
- [51] J. J. Bird, D. R. Faria, L. J. Manso, A. Ekárt, and C. D. Buckingham, "A deep evolutionary approach to bioinspired classifier optimisation for brain-machine interaction," *Complexity*, vol. 2019, pp. 1–14, Mar. 2019.
- [52] J. J. Bird, A. Ekart, C. Buckingham, and D. R. Faria, "Mental emotional sentiment classification with an EEG-based brain-machine interface," in *Proc. Int. Conf. Digit. Image Signal Process. (DISP)*, 2019, pp. 1–8.
- [53] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [54] E. Calore and S. F. Schifano, "Performance assessment of FPGAs as HPC accelerators using the FPGA empirical roofline," in *Proc. 31st Int. Conf. Field-Program. Log. Appl. (FPL)*, Aug. 2021, pp. 83–90.



AHMED K. JAMEIL received the bachelor's degree in electrical and electronics engineering from the University of Technology, Iraq, in 2001, and the M.Sc. degree in computer and microelectronic systems engineering from the University of Technology Malaysia (UTM). He is currently pursuing the Ph.D. degree with Brunel University London, U.K. He worked as an Engineer with the University of Technology, from 2003 to 2006. Following that, he was worked with the College of Engineering, University of Diyala, Iraq, as an Engineer, from 2006 to 2010. He is a member of Staff with the Department of Computer Engineering, College of Engineering, University of Diyala. He has published several journal and conference papers. His research interests include advanced computer architecture, VLSI circuits and design, and integrated circuit testing. In his current study, he is interested in designing digital integrated circuits and processor for artificial intelligence.



HAMED AL-RAWESHIDY (Senior Member, IEEE) received the B.Eng. and M.Sc. degrees from the University of Technology, Baghdad, Iraq, in 1977 and 1980, respectively, the postgraduate Diploma degree from the University of Glasgow, Glasgow, U.K., in 1987, and the Ph.D. degree from the University of Strathclyde, Glasgow, in 1991. He was with the Space and Astronomy Research Centre, Baghdad; PerkinElmer, Waltham, MA, USA; British Telecom, London, U.K.; the University of Oxford, Oxford, U.K.; Manchester Metropolitan University, Manchester, U.K.; and the University of Kent, Canterbury, U.K. He is currently the Director of the Wireless Network Communications Centre, Brunel University London.

• • •