

Efficient ECG-based Atrial Fibrillation Detection via Parameterised Hypercomplex Neural Networks

Leonie Basso
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
basso@l3s.de

Zhao Ren
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
zren@l3s.de

Wolfgang Nejdl
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
nejdl@l3s.de

Abstract—Atrial fibrillation (AF) is the most common cardiac arrhythmia and associated with a high risk for serious conditions like stroke. The use of wearable devices embedded with automatic and timely AF assessment from electrocardiograms (ECGs) has shown to be promising in preventing life-threatening situations. Although deep neural networks have demonstrated superiority in model performance, their use on wearable devices is limited by the trade-off between model performance and complexity. In this work, we propose to use lightweight convolutional neural networks (CNNs) with parameterised hypercomplex (PH) layers for AF detection based on ECGs. The proposed approach trains small-scale CNNs, thus overcoming the limited computing resources on wearable devices. We show comparable performance to corresponding real-valued CNNs on two publicly available ECG datasets using significantly fewer model parameters. PH models are more flexible than other hypercomplex neural networks and can operate on any number of input ECG leads.

Index Terms—Atrial fibrillation, ECG analysis, hypercomplex domain, lightweight neural networks

I. INTRODUCTION

Atrial fibrillation (AF), characterised by an irregular beating of the atrial chambers of the heart, is the most common serious abnormal heart rhythm and a risk factor for strokes [1]. A timely diagnosis of AF is crucial for the patient’s health. The most reliable way to test for cardiac arrhythmia is to analyse the electrical activity of a heart as electrocardiograms (ECGs) [2]. However, diagnosis of AF from ECGs requires well-trained medical professionals, and AF often remains unrecognised during conventional short-time monitoring due to its episodic nature in early stages [2]. In the era of Internet-of-Medical-Things, wearable devices such as Holter monitors and smartwatches are enabling the diagnosis of cardiac diseases in daily life based on long-interval ECG signals [2], [3]. A manual diagnosis from long-term ECG signals by medical professionals is time-consuming [2], which results in a bottleneck for self-care of AF. Therefore, automated analysis of long-interval ECG signals is essential for AF detection. Recently, deep neural networks (DNNs) have achieved success in ECG signal analysis [4], such as arrhythmia detection [5] and AF classification [6]. However, the complexity of large

DNNs limits their deployment on wearable devices [4], which motivates the need for small and efficient models.

A variety of model compression techniques have been proposed. For example, pruning aims to discard unnecessary network connections and quantisation targets to represent model weights with less bits [7]. Knowledge distillation builds a shallow student model trained with a pre-trained deep teacher model [8], [9]. In contrast to real-valued neural networks used in the above approaches, quaternion neural networks (QNNs) build lightweight neural networks by inherently changing the construction of layer weights. The input elements are processed as entities of one real and three imaginary components based on 4-dimensional hypercomplex quaternion numbers [10]. QNNs only need 1/4 learnable parameters compared to corresponding real-valued neural networks and can capture internal latent relations between input channels [10]. Nevertheless, their dimensionality is limited because operations in hypercomplex space are only predefined in limited dimensions, such as 4D, 8D, and 16D.

Parameterised hypercomplex (PH) neural networks [11], [12] were proposed to overcome the dimensionality restrictions of QNNs. They can operate in any input dimension, allowing a flexible and domain-specific processing of multidimensional inputs. PH layers use Kronecker product properties to construct weight matrices that reduce the number of learnable parameters to $1/n$ compared to corresponding real-valued neural networks based on the hyperparameter n . Similar to QNNs, PH models are also able to learn internal relations between channels. Specifically, PH multiplication (PHM) [11] has been proposed to replace real-valued matrix multiplication in fully-connected (FC) layers. PH convolutional (PHC) layers [12] extend PHM to convolution and facilitate the development of efficient deep convolutional neural networks (CNNs). Therefore, PH models can help apply deep learning to AF detection on wearable devices by using significantly fewer parameters than real-valued models.

To the best of the authors’ knowledge, hypercomplex DNNs have rarely been applied to ECG signals. The study in [13] used QNNs to model cardiac velocity patterns from ECG-related vectorcardiogram signals, which record spatio-temporal information of cardiac electrical activity. The 4D quaternions are suitable to model rotations in a 3D space,

This research was funded by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003.

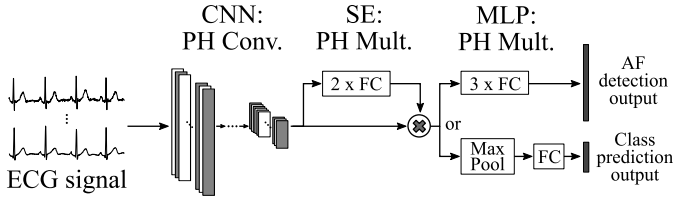


Fig. 1: Proposed PH-CNN architecture, including three modules: (1) a CNN, (2) a squeeze-and excitation (SE) attention, and (3) a multilayer perceptron (MLP) classifier. Compared to real-valued DNNs, parameterised hypercomplex (PH) convolution and multiplication replace real-valued convolutional and fully-connected (FC) layers, respectively. We construct separate models for two tasks: (a) AF detection, where every sampling point of the input ECG signal gets classified as AF/non-AF, and (b) global abnormality classification, where the output is a vector of class probabilities.

however, not applicable to ECG signals with different numbers of leads. In this work, we propose lightweight PH-CNNs with PHM and PHC layers for AF detection and abnormality classification from ECG signals. Our contributions are manifold. (1) PH-CNNs can significantly reduce the number of model parameters and achieve comparable performance to corresponding real-valued models, enabling applications of effective AF detection on wearable devices. (2) PH-CNNs are flexible in input dimensions and can effectively process ECG signals with different numbers of leads. (3) Learning inter-channel relations between the different ECG leads with PH-CNNs is promising to improve the model performance. We conduct experiments to demonstrate the performance and efficiency of the approach with several CNN architectures and channel dimensionalities on two publicly available datasets.

II. METHODOLOGY

We propose a PH-CNN architecture that takes raw ECG signals as input and consists of three modules (see Fig. 1): (1) A CNN with PHC layers extracts features, (2) a squeeze-and-excitation (SE) [14] attention mechanism with PHM improves channel interdependencies, and (3) a multilayer perceptron (MLP) classifier with PHM produces predictions. We construct separate models for two objectives: (1) *AF detection* and (2) *rhythm/morphology abnormality classification*. AF detection aims to detect the on- and offsets of AF episodes by predicting AF or non-AF for each sampling point, which helps in developing personalised treatments [15]. The classification of abnormalities, including AF, is performed at the global level for each ECG recording.

A. Parameterised Hypercomplex Neural Network

A real-valued FC layer has a weight matrix $\mathbf{W}_{\text{FC}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ for in- and output dimensions d_{in} and d_{out} . Thus, the parameter size of a real-valued FC layer is $\mathcal{O}(d_{\text{out}} \cdot d_{\text{in}})$. The parameter size of a real-valued 1D convolutional layer is $\mathcal{O}(d'_{\text{out}} \cdot d'_{\text{in}} \cdot k)$ for a weight matrix $\mathbf{W}_{\text{Conv}} \in \mathbb{R}^{d'_{\text{out}} \times d'_{\text{in}} \times k}$, in- and output dimensions d'_{in} and d'_{out} , and the size k of a 1D kernel.

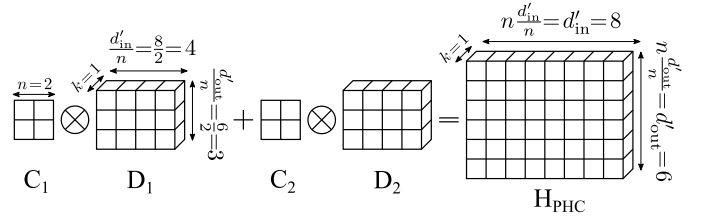


Fig. 2: Example of a weight matrix \mathbf{H}_{PHC} in a parameterised hypercomplex convolutional (PHC) layer [12] when $n = 2$. \mathbf{H}_{PHC} is built by the sum of the Kronecker products of matrices $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n \times n}$ and $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{(d'_{\text{out}}/n) \times (d'_{\text{in}}/n) \times k}$. The row or column size of a Kronecker product is the multiplication of the row or column size of the input matrices, respectively, therefore $\mathbf{H}_{\text{PHC}} \in \mathbb{R}^{d'_{\text{out}} \times d'_{\text{in}} \times k}$. In this example, the input channel size $d_{\text{in}} = 8$, output size $d_{\text{out}} = 6$, and kernel size $k = 1$. (Adapted from [11])

PHM [11] and PHC [12] layers generalise hypercomplex operations and can operate in arbitrary domains $n\mathbb{D}$, where the dimensionality $n \in \mathbb{N}$ can be chosen depending on the input dimensionality or adapted as a hyperparameter. The input elements are processed as hypercomplex numbers of dimension n . Due to weight sharing over the n dimensions, PH models are able to capture internal relations between channels.

In a PHM layer [11], the multiplication of an input $x \in \mathbb{R}^{d_{\text{in}}}$ or $x \in \mathbb{R}^{d_{\text{in}} \times t}$ with a weight matrix $\mathbf{H}_{\text{PHM}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ to produce an output $y \in \mathbb{R}^{d_{\text{out}}}$ or $y \in \mathbb{R}^{d_{\text{out}} \times t}$ is defined as

$$y = \text{PHM}(x) = \mathbf{H}_{\text{PHM}} \cdot x + \mathbf{b}_{\text{FC}}. \quad (1)$$

\mathbf{H}_{PHM} is built by the sum of n Kronecker products of learnable matrices $\mathbf{A}_i \in \mathbb{R}^{n \times n}$ and $\mathbf{B}_i \in \mathbb{R}^{(d_{\text{out}}/n) \times (d_{\text{in}}/n)}$ for $i = 1, \dots, n$,

$$\mathbf{H}_{\text{PHM}} = \sum_{i=1}^n \mathbf{A}_i \otimes \mathbf{B}_i. \quad (2)$$

The hypercomplex multiplication of the weight parameters \mathbf{B} and the input x relies on interactions between imaginary units, which are modelled by the algebra rules in \mathbf{A}_i . These define arithmetic operations, such as multiplication, in the $n\mathbb{D}$ hyperspace and determine the arrangement of \mathbf{B}_i in the final weight matrix \mathbf{H}_{PHM} . PH layers learn the algebra rules as matrices \mathbf{A}_i directly from the data. Thus, they don't rely on predefined rules, which exist, for example, for the complex (i.e., $2\mathbb{D}$) and quaternion (i.e., $4\mathbb{D}$) domains but not for $6\mathbb{D}$.

In a PHC layer [12], the convolution of an input $x' \in \mathbb{R}^{d'_{\text{in}} \times t_{\text{in}}}$ with a weight matrix \mathbf{H}_{PHC} to produce an output $y' \in \mathbb{R}^{d'_{\text{out}} \times t_{\text{out}}}$ is defined as

$$y' = \text{PHC}(x') = \mathbf{H}_{\text{PHC}} * x' + \mathbf{b}_{\text{Conv}}. \quad (3)$$

Analogous to \mathbf{H}_{PHM} , $\mathbf{H}_{\text{PHC}} \in \mathbb{R}^{d'_{\text{out}} \times d'_{\text{in}} \times k}$ is built by the sum of n Kronecker products of learnable matrices $\mathbf{C}_i \in \mathbb{R}^{n \times n}$ and $\mathbf{D}_i \in \mathbb{R}^{(d'_{\text{out}}/n) \times (d'_{\text{in}}/n) \times k}$ for a 1D kernel with size k (see Fig 2). \mathbf{D}_i represents the i -th collection of convolution filters that are organised after the algebra rules \mathbf{C}_i ,

$$\mathbf{H}_{\text{PHC}} = \sum_{i=1}^n \mathbf{C}_i \otimes \mathbf{D}_i. \quad (4)$$

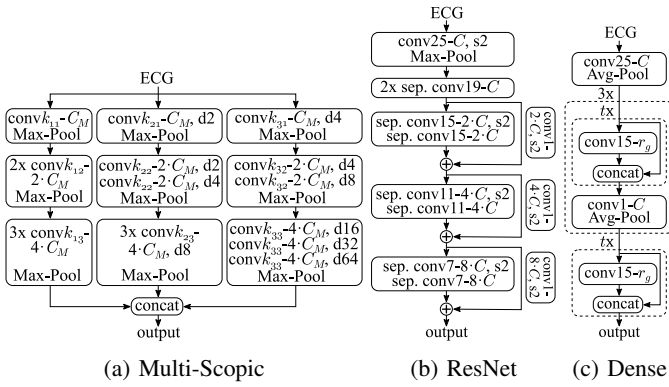


Fig. 3: Configurations of the three CNN backbones. Convolutional layers are denoted as “conv(kernel size-number of channels)” with “s(stride size)” and “d(dilation rate)” (default 1). The output is then fed into an attention module (see Fig. 1). The selected parameter values depend on the input dimensionality: for 2-channel input $C_M=16$, $C=64$, $t=4$, growth rate $r_g=16$, $k=[3, 3, 3]$, $[5, 5, 3]$, $[9, 7, 5]$, and for 12-channel input $C_M=24$, $C=72$, $t=6$, $r_g=12$, $k=[3 \times [11, 7, 5]]$.

The parameter size of a PHM layer can be approximated to $\mathcal{O}(d_{\text{out}} \cdot d_{\text{in}}/n)$ and of a PHC layer to $\mathcal{O}(d'_{\text{out}} \cdot d'_{\text{in}} \cdot k/n)$, if $d_{\text{out}} \cdot d_{\text{in}} \gtrsim n^4$ holds. This assumption is mild for real-world problems, where usually large numbers of filters are employed. This leads to a parameter reduction to approximately $1/n$ compared to a standard real-valued FC or convolutional layer, respectively. Note that the dimensions d_{in} , d_{out} , d'_{in} and d'_{out} have to be divisible by n .

B. Model architectures

We select three lightweight CNN backbones: Multi-Scopic CNN [16], ResNet [17], and DenseNet [18]. The Multi-Scopic CNN has been used as a default model for an AF detection benchmark [19] and both ResNets and DenseNets have been successfully applied for arrhythmia classification from ECGs [5], [19]. The CNN model configurations are shown in Fig. 3. The Multi-Scopic CNN consists of three parallel CNN branches that use different dilation rates to combine information from different receptive fields. The DenseNet architecture consists of several dense blocks, which concatenate output feature maps of a layer with input feature maps and are connected by transition blocks. In ResNets, the output of a residual block is added to the input that is upsampled in a shortcut connection. To reduce model parameters, the residual blocks here use separable convolutions, which consist of a depth-wise and a point-wise operation. In depth-wise convolutions, the input channels are processed separately with $d'_{\text{in}}=1$. Therefore, PHC is not used for depth-wise convolution, but for the 1×1 point-wise convolutions in separable convolutions and in the shortcut connection of each residual block.

Each CNN module is followed by an SE block [14], which performs feature recalibration by multiplying input feature maps with channel weights learnt by two FC layers. The MLP classifier for AF detection consists of three FC layers and

TABLE I: Distribution of classes in the training, test and validation sets for segments from the CPSC 2021 and CPSC 2018 datasets.

#	CPSC 2021				CPSC 2018						
	non-AF	pers. AF	PAF	Σ	NSR	AF	I-AVB	LBBB	RBBB	PAC	Σ
Train	35,643	12,442	6,169	54,254	374	508	290	96	729	232	2229
Val.	23,146	14,382	1,947	39,475	261	358	225	66	559	192	1661
Test	17,244	11,317	3,153	31,714	283	354	206	73	566	190	1672

a linear interpolation to recover the sequence length of the inputs. For abnormality classification, a probability for each class is produced by max pooling and an FC layer (see Fig. 1).

III. EXPERIMENTS AND RESULTS

A. Datasets

For this study, we use two datasets with different sources, number of recorded ECG leads, and tasks.

CPSC 2021. We use 2-lead data from the CPSC 2021 dataset [20] for the detection of paroxysmal atrial fibrillation (PAF) events, which are intermittent AF episodes that end within less than a week. The data include signals from lead I and lead II of long-term dynamic ECGs recorded with wearable Holter monitors at 200 Hz. The publicly available training set consists of 1,436 records from 23 PAF, 24 persistent AF, and 53 non-AF patients. Due to the unreleased test set, we perform a random subject-independent split in 40% train, 30% validation, and 30% test sets. All classes appear in all subsets with approximately the same distribution as in the entire dataset (see Table I). Apart from AF detection, the global rhythm is derived from the AF label of each sampling point in order to perform a three-class classification: *non-AF* (no AF rhythm present), *PAF* (both AF and non-AF episodes), and *persistent AF* (continuous AF). During preprocessing, z-score normalisation and a bandpass filter with cutoff frequencies of 0.5 and 45 Hz were applied, following [21]. The preprocessed signals were then sliced into 30 s segments with 15 s overlap.

CPSC 2018. We use 12-lead ECG signals from the CPSC 2018 dataset [22] as provided in the PhysioNet/Computing in Cardiology Challenge (CinC) 2020 [23], [24]. The dataset includes 5,271 records that are collected from in-hospital patients at 500 Hz. Again, we split the data subject-independently into 40% train, 30% validation and 30% test sets considering class balance (see Table I). A multi-label classification task is performed with the following rhythm/morphology classes: normal sinus rhythm (NSR), AF, first-degree atrioventricular block (I-AVB), left bundle branch block (LBBB), right bundle branch block (RBBB), and premature atrial contraction (PAC). The signals are z-score normalised and a 10 s segment is extracted from the middle of each record.

B. Experimental Setup

For CPSC 2021, we run experiments with $n=2$ in the PH models, matching the number of ECG leads, and also with $n=4$, where the input signals are zero-padded to 4 channels. For CPSC 2018, we test $n=2, 4, 6, 12$ to investigate a trade-off between accuracy and model complexity. The value of n is

the same for all layers except for the final MLP layers, where $n=1$ because d_{out} must be divisible by n .

Direct comparison with previous studies is difficult due to different data sets and splits. We compare the performance of our framework with two popular deep ResNets for arrhythmia classification [5], [25], and a Multi-Scopic CNN with SE attention and an additional bidirectional long short-term memory (LSTM) [21], following the benchmarks proposed in [19].

C. Evaluation Metrics

We evaluate the classification performance based on the unweighted average recall (UAR) to account for class imbalance. For the classification of classes C , the UAR is based on true-positive (TP) and false-negative (FN) samples,

$$\text{UAR} = \frac{1}{|C|} \sum_{c \in C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}. \quad (5)$$

To assess the significance of the comparison between UARs of PH-CNNs and real-valued networks, we perform a one-tailed z-test on the test sets. In addition, the challenge score S_{AF} of CPSC 2021 [20] is used to evaluate the correct detection of AF events. S_{AF} evaluates the overall rhythm prediction as well as the detection of start and end points of an AF episode.

D. Model Training

The models are implemented using the `torch_ecg` package [26], [19]. We generally follow the training setup proposed in their CPSC 2021 and CinC 2020 (for CPSC 2018) benchmark tests. We use the Asymmetric Loss [27] to handle the class imbalance, a learning rate of 0.0001 and the AMSGrad variant of the AdamW optimiser [28]. The best model is determined based on binary-cross-entropy (for AF detection) or global accuracy (for abnormality classification) on the validation set. The batch size is set to 64. The code is available at: <https://github.com/leibniz-future-lab/Hypercomplex-ECG>.

E. Results and Discussion

AF Detection. Using PH-Multi-Scopic as the CNN backbone yielded the highest score S_{AF} for detecting the start and end points of AF episodes on CPSC 2021 data (see Table II). This value is also comparable to most challenge submissions [20], however, a direct comparison is difficult due to different test sets. PH-Multi-Scopic ($n=2$, $p<0.001$ in one-tailed z-test) and PH-ResNet ($n=4$, $p<0.05$) models achieve a higher UAR than the corresponding real-valued networks. A possible explanation is the ability of PH layers to learn inter-channel relations due to weight sharing. These two models still achieve high UARs when the model is further compressed with $n=4$. However, the UAR decreases for PH-DenseNet ($n=2$, $p<0.001$), while S_{AF} is comparable. Both metrics are calculated from the same sequence output, but the UAR weights all classes equally and doesn't evaluate the AF on-/offsets. The PH-DenseNets generalise worse on the PAF class in the validation and test sets, possibly due to DenseNet's simpler structure and limited ability to learn useful features from this data. The previously proposed models [5], [25],

TABLE II: Comparison between parameterised hypercomplex (PH) models with hyperparameter n and real-valued models (n is '-') for AF detection on the 2-lead ECG data in CPSC 2021. Results are reported for Unweighted Average Recall (UAR) and Score S_{AF} from CPSC 2021 on the validation and test sets. C.Par. = CNN Parameters, Multi-Sc. = Multi-Scopic.

Network	n	#C.Par.	#Params	UAR in % \uparrow		S_{AF} \uparrow	
				Val	Test	Val	Test
ResNet [5]	-	26,499k	26,746k	87.40	83.86	1.3719	1.3804
ResNet [25]	-	8,708k	9,463k	81.24	82.04	1.3389	1.3622
Multi-Sc.+LSTM[21]	-	138k	2,850k	84.16	87.14	1.3917	1.4258
Multi-Scopic	-	138k	213k	81.25	83.51	1.8944	2.0486
PH-Multi-Scopic	2	71k	112k	82.06	86.06	1.6170	2.0560
PH-Multi-Scopic	4	38k	61k	88.79	82.62	1.6989	2.4488
ResNet	-	720k	934k	80.14	86.17	1.3392	1.4622
PH-ResNet	2	373k	497k	86.22	81.88	1.3721	1.4424
PH-ResNet	4	200k	279k	88.61	86.68	1.4198	1.4629
DenseNet	-	369k	423k	87.29	83.49	1.4065	1.3636
PH-DenseNet	2	187k	215k	83.44	78.80	1.3523	1.3672
PH-DenseNet	4	98k	113k	76.26	78.72	1.3490	1.3431

[21] achieve comparable UARs to the best proposed small-scale model, while using significantly more parameters. They also have a notably lower S_{AF} than PH-Multi-Scopic and are therefore worse at detecting the start and end of AF episodes.

Abnormality Classification. The classification performance of PH and real-valued models on the 12-lead CPSC 2018 data is comparable for all three CNN backbones (see Table III). The highest UAR overall is achieved by PH-Multi-Scopic models. An increased n doesn't lead to a declining performance, therefore, we observe no large trade-off between increased parameter saving and classification performance in this setting. This could again be explained with the ability of PH models to capture the relations between different channels, especially for a larger n that covers all input channels. Again, the previously proposed models [5], [25], [21] don't exceed the classification performance of the small-scale models while using significantly more parameters ($p<0.05$ for PH-Multi-Scopic $n=12$ and ResNet [5]).

Model compression. A large parameter reduction can be observed using PHC layers in all CNNs, saving already approximately -50% CNN parameters for $n=2$. The number of parameters is further reduced as n increases, leading to a substantial model compression when $n=12$ on 12-lead data, where the number of parameters in all three CNNs is reduced by more than 80%. Even though the depth-wise convolutions in ResNet couldn't be adapted to PHC, the number of parameters are notably reduced due to PH point-wise convolutions. The CNN module accounts for the vast majority of the model size, so its compression has the greatest impact. PHM layers also reduce the parameter size of the FC layers in the SE and MLP module to approximately $1/n$, resulting in a large overall parameter reduction. The value of n can be chosen flexibly during model creation based on the data dimensionality and desired model compression.

TABLE III: Comparison between parameterised hypercomplex (PH) models with hyperparameter n and real-valued models (n is ‘-’) for abnormality classification on 12-lead ECG data from CPSC 2018. Results are reported for Unweighted Average Recall (UAR) on the validation and test sets.

Network	n	#CNN Params	#Params	UAR in % \uparrow	
				Val	Test
ResNet [5]	-	26,511k	26,514k	92.43	90.34
ResNet [25]	-	8,719k	8,726k	91.71	89.09
Multi-Sc.+LSTM [21]	-	430k	2,135k	92.43	89.99
Multi-Scopic	-	430k	473k	90.24	90.68
PH-Multi-Scopic	2	219k (-49%)	252k	91.53	88.64
PH-Multi-Scopic	4	113k (-74%)	141k	91.59	89.65
PH-Multi-Scopic	6	80k (-81%)	106k	93.14	89.59
PH-Multi-Scopic	12	71k (-83%)	98k	90.55	92.03
ResNet	-	926k	1,096k	91.02	90.20
PH-ResNet	2	478k (-48%)	607k	91.50	89.88
PH-ResNet	4	253k (-73%)	361k	90.89	90.91
PH-ResNet	6	180k (-81%)	281k	91.95	90.77
PH-ResNet	12	122k (-87%)	173k	90.82	89.78
DenseNet	-	499k	511k	90.03	89.10
PH-DenseNet	2	253k (-49%)	262k	92.24	89.85
PH-DenseNet	4	132k (-74%)	139k	90.56	87.90
PH-DenseNet	6	95k (-81%)	102k	91.16	89.32
PH-DenseNet	12	96k (-80%)	104k	92.68	90.96

IV. CONCLUSION AND OUTLOOK

We proposed a lightweight PH neural network framework for AF detection from ECG signals based on PH operations for multiplication and convolution. We demonstrated comparable performance to corresponding real-valued networks for different CNN backbones on two datasets, while significantly reducing the parameter size. The dimensionality n can be chosen flexibly based on the data or tuned for further compression. In future work, PH architectures can be combined with other advanced model compression techniques, such as knowledge distillation [8], to further reduce the model size. Implementation on wearable devices is further limited by the need for low computational complexity and power consumption, which should also be investigated in the future. PH layers can also benefit the automatic analysis of other data types from wearable devices, such as photoplethysmography (PPG).

REFERENCES

- [1] G. Lippi, F. Sanchis-Gomar, and G. Cervellin, “Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge,” *Int. J. Stroke*, vol. 16, pp. 217–221, 2021.
- [2] A. Rizwan, A. Zoha, I. B. Mabrouk, H. M. Sabbour, A. S. Al-Sumaiti, A. Alomainy, M. A. Imran, and Q. H. Abbasi, “A review on the state of the art in atrial fibrillation detection enabled by machine learning,” *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 219–239, 2021.
- [3] J. K. Sajeev, A. N. Koshy, and A. W. Teh, “Wearable devices for cardiac arrhythmia detection: A new contender?” *Intern. Med. J.*, vol. 49, no. 5, pp. 570–573, 2019.
- [4] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, “Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review,” *Comput. Biol. Med.*, vol. 122, p. 103801, 2020.
- [5] 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,” *Nat. Med.*, vol. 25, pp. 65–69, 2019.

- [6] R. S. Andersen, A. Peimankar, and S. Puthusserypady, “A deep learning approach for real-time detection of atrial fibrillation,” *Expert Syst. Appl.*, vol. 115, pp. 465–473, 2019.
- [7] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding,” in *Proc. ICLR*, 2015.
- [8] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [9] A. Polino, R. Pascanu, and D. Alistarh, “Model compression via distillation and quantization,” in *Proc. ICLR*, 2018, pp. 1–21.
- [10] T. Parcollet, M. Morchid, and G. Linares, “A survey of quaternion neural networks,” *Artif. Intell. Rev.*, vol. 53, pp. 2957–2982, 2020.
- [11] A. Zhang, Y. Tay, S. Zhang, A. Chan, A. T. Luu, S. C. Hui, and J. Fu, “Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters,” in *Proc. ICLR*, 2021, pp. 1–13.
- [12] E. Grassucci, A. Zhang, and D. Comminiello, “PHNNs: Lightweight neural networks via parameterized hypercomplex convolutions,” *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–13, 2022.
- [13] P. D. Cruces, R. Correa, E. Laciár, and P. Arini, “Quaternion neural network with temporal feedback calculation: Application to cardiac vector velocity during myocardial infarction,” *Revista Argentina de Bioingeniería*, vol. 22, no. 3, pp. 60–64, 2018.
- [14] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. CVPR*, 2018, pp. 7132–7141.
- [15] R. R. De With, Ö. Erkiner, M. Rienstra, B.-O. Nguyen, F. W. Körver, D. Linz, H. Cate Ten, H. Spronk, A. A. Kroon, A. H. Maass *et al.*, “Temporal patterns and short-term progression of paroxysmal atrial fibrillation: Data from RACE V,” *EP Europace*, vol. 22, no. 8, pp. 1162–1172, 2020.
- [16] W. Cai and D. Hu, “QRS complex detection using novel deep learning neural networks,” *IEEE Access*, vol. 8, pp. 97082–97089, 2020.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.
- [18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. CVPR*, 2017, pp. 4700–4708.
- [19] H. Wen and J. Kang, “A novel deep learning package for electrocardiography research,” *Physiol. Meas.*, vol. 43, no. 11, p. 115006, 2022.
- [20] X. Wang, C. Ma, X. Zhang, H. Gao, G. D. Clifford, and C. Liu, “Paroxysmal atrial fibrillation events detection from dynamic ECG recordings: The 4th China Physiological Signal Challenge 2021,” 2021. [Online]. Available: <http://2021.icbeb.org/CPSC2021>
- [21] H. Wen and J. Kang, “A comparative study on neural networks for paroxysmal atrial fibrillation events detection from electrocardiography,” *J. Electrocardiol.*, vol. 75, pp. 19–27, 2022.
- [22] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He *et al.*, “An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection,” *J. Med. Imaging Health Inf.*, vol. 8, pp. 1368–1373, 2018.
- [23] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A. K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi *et al.*, “Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020,” *Physiol. Meas.*, vol. 41, p. 124003, 2020. [Online]. Available: <https://moody-challenge.physionet.org/2020/>
- [24] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [25] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr. *et al.*, “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nat. Commun.*, vol. 11, no. 1, p. 1760, 2020.
- [26] H. Wen and J. Kang, “torch_ecg: An ECG deep learning framework implemented using PyTorch,” 2022, v0.0.26. [Online]. Available: https://github.com/DeepPSP/torch_ecg
- [27] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, “Asymmetric loss for multi-label classification,” in *Proc. ICCV*, 2021, pp. 82–91.
- [28] S. J. Reddi, S. Kale, and S. Kumar, “On the convergence of adam and beyond,” in *Proc. ICLR*, 2018, pp. 1–23.