

Towards Hardware Supported Domain Generalization in DNN-based Edge Computing Devices for Health Monitoring

Johnson Loh, Lyubov Dudchenko, Justus Viga and Tobias Gemmeke, *Senior Member, IEEE*

Abstract—Deep neural network (DNN) models have shown remarkable success in many real-world scenarios, such as object detection and classification. Unfortunately, these models are not yet widely adopted in health monitoring due to exceptionally high requirements for model robustness and deployment in highly resource-constrained devices. In particular, the acquisition of biosignals, such as electrocardiogram (ECG), is subject to large variations between training and deployment, necessitating domain generalization (DG) for robust classification quality across sensors and patients. The continuous monitoring of ECG also requires the execution of DNN models in convenient wearable devices, which is achieved by specialized ECG accelerators with small form factor and ultra-low power consumption. However, combining DG capabilities with ECG accelerators remains a challenge. This article provides a comprehensive overview of ECG accelerators and DG methods and discusses the implication of the combination of both domains, such that multi-domain ECG monitoring is enabled with emerging algorithm-hardware co-optimized systems. Within this context, an approach based on correction layers is proposed to deploy DG capabilities on the edge. Here, the DNN fine-tuning for unknown domains is limited to a single layer, while the remaining DNN model remains unmodified. Thus, computational complexity (CC) for DG is reduced with minimal memory overhead compared to conventional fine-tuning of the whole DNN model. The DNN model-dependent CC is reduced by more than $2.5 \times$ compared to DNN fine-tuning at an average increase of F1 score by more than 20 % on the generalized target domain. In summary, this article provides a novel perspective on robust DNN classification on the edge for health monitoring applications.

Index Terms—ECG processing, correction layer, domain generalization, domain shift, hardware accelerator

I. INTRODUCTION

As wearable devices increasingly become available in people's daily lives, there is a growing need for edge computing devices capable of analysing personal data in a privacy preserving system. Complex machine learning models fitted to the edge enable wearable devices with advanced processing capabilities without transmission of personal data to third-party systems. Deep neural networks (DNNs) have shown impressive results in several applications, such as image

The authors are with the Chair of Integrated Digital Systems and Circuit Design, RWTH Aachen University, 52074 Aachen, Germany (e-mail: loh@ids.rwth-aachen.de; gemmeke@ids.rwth-aachen.de).

This work is partially funded by the German Federal Ministry of Education and Research under grant no. 03ZU1106CA (Clusters4Future – NeuroSys) and partially by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection under grant no. 67KI32006A (RESCALE).

classification [1] and speech recognition [2]. Nevertheless, their high computational complexity (CC) is still a challenge for deployment in resource-constrained edge devices. To address this challenge, DNN accelerators have been developed to enable the efficient execution of large DNN models on specialized hardware [3]. Although current DNN accelerators can perform inference tasks efficiently, in practice, the features in the actual data deviate from those in the data used during training - the *domain shift* problem. In practical settings, the domain shift is caused by differences in the background environment, measurement setup, or differences in the measured subject.

The generalization to these out-of-distribution (OOD) data is well addressed for a wide range of applications [4], [5], in particular image processing tasks [6], [7]. Figure 1 shows a brief summary of the problem of OOD generalization. While classical DNN training assumes independent and identically distributed (IID) data, i.e. source domain (SD) and target domain (TD) are similarly distributed, the OOD generalization problem introduces a discrepancy between SD and TD. Typically, data from SD are utilized to train the DNN during a training phase, while data from TD is used to evaluate the quality of generalization during a testing phase. Dependent on the availability of labels in SD/TD and data during training, the general field of OOD generalization can be further subdivided into e.g. detection of new classes (Zero-Shot Learning [8]) or transfer of features to new classification tasks (Transfer Learning [9]). Here, domain generalization (DG) deals with the increased robustness against domain shift, while the classification task remains the same as during training. Especially, working with faint features as in electrocardiogram (ECG) signals requires generalizing properties as domain shift heavily affects the performance of pre-trained DNN models [10].

In this case the challenge is the support of DG in wearables, which go beyond DNN parameter reconfiguration or on-line DNN training as supported in state-of-the-art ECG accelerators. To the best of our knowledge, there is no comprehensive work addressing the acceleration of DG-based algorithms in hardware (HW) and, hence, this field poses a new and promising direction for future DNN systems. The application case of ECG monitoring is well suited for an initial feasibility study, as large domain variations exist between collected databases and actual deployment on wearables. Further, on-device processing is necessary to ensure that personal data is handled locally.

Therefore, this work summarizes DG methods and analyses them with the intention for deployment on dedicated acceler-

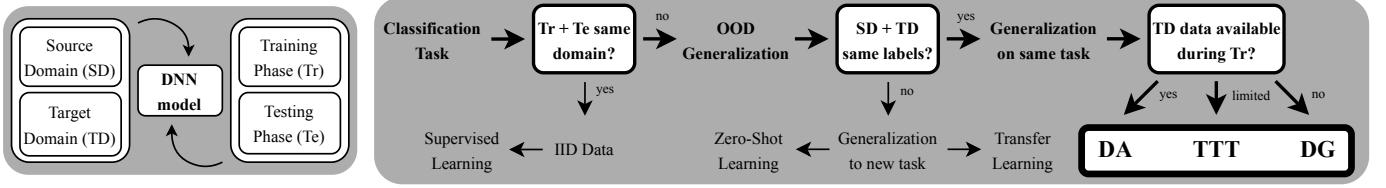


Fig. 1: Problem Overview

ators in edge devices. First, the state-of-the-art is reviewed to compile common features and trends for DG in general and specific to ECG signals. Then, the deployment scenarios are outlined to investigate the impact on HW resources and communication (see Section II). As an initial entry point for DG specific accelerators, the concept of correction layers (CLs) are introduced as a means for DG with minimal HW overhead (see Section III). For atrial fibrillation classification, we show statistically robust classification performance improvements, while the original model is unmodified and augmented with a single layer. The evaluation of memory and computational complexity shows that CL outperforms conventional DNN fine-tuning, while quantitative gains are dependent on DNN architecture and CL position (Section IV). This minimally invasive approach serves as a baseline for further improvements on DG implementations through algorithm-hardware co-optimizations.

The main contributions of this work are summarized as follows:

- Review of DG methods is provided in the context of ECG classification for deployment on dedicated HW accelerators
- Correction layers (CLs) are proposed in a case study to perform DG with minimal modifications on pre-trained models and robust classification performance
- Evaluation of CL performance during inference using implemented HW accelerators synthesized in 22 nm CMOS
- Estimation of memory and computational complexity of CL during training

II. BACKGROUND

As indicated in Section I, the spectrum of DG methods is large, while only a subset is currently applied on ECG classification in state-of-the-art literature. Hence, this section briefly introduces existing state-of-the-art methods and analyses constraints in their applicability to the ECG domain and for HW acceleration.

A. DG method overview

As depicted in Fig. 1, OOD generalization covers multiple application fields, while the generalization for the sake of model robustness, i.e. classification of same classes in SD and TD, also differentiates between domain adaptation (DA), test-time training (TTT) and DG. The main difference is the availability of TD data during the training process. The extreme cases assume the corner scenarios, where full [11]

TABLE I: Required component modifications during DNN training for DG algorithms

	Input data	Labels	Addtl. Models	DNN Arch.	DNN param.	optim. function
Domain Alignment	✗	✗	✗	N/A	✓	✓
Meta-Learning	N/A	N/A	✓	N/A	✓	✓
Ensemble Learning	✗	✗	✗	✓	✓	N/A
Data Augmentation	✓	✓	✗	✗	✓	✗
Self-Supervised Learning	✓	✓	✗	✗	✓	N/A
Disentangled Representation	✗	✗	✗	N/A	✓	✓
Regularization	✗	✗	✗	✗	✓	✓
Expert knowledge	✗	✗	✓	N/A	✗	✗

or no information [12] in the TD is available, respectively. Adaptations during test-time, however, uses limited data from TD to fine-tune the model during deployment [13]. As the target of this work is to increase ECG classification robustness, we look at both corner cases and refer to both as DG for the sake of simplicity.

Table I summarizes recent DG methods. Starting from a trained DNN model, modifications on the reference are performed. Some approaches, such as data augmentation and self-supervised learning, aim to extend the training dataset with additional synthetic [14] or unlabeled samples [15], which enclose the features of TD. Other approaches, such as meta-learning, require additional models to incorporate domain shift in the training process [16]. Modifications on the original DNN model architecture are also utilized, e.g. with ensembles of domain specific classifiers [17]. Another common method is to modify the training algorithm, e.g. by including additional loss terms to achieve domain-invariant features [18]–[20]. In general, most methods require an update of DNN model parameters to adjust for the new TD. In the perspective of DG HW acceleration, methods will be preferred, which require least modification on the original inference model to calculate the DG algorithm. Here, DG complexity mainly depends on the target application.

B. DG applied on ECG

In this work, we investigate ECG classification as an example application. Figure 2 summarizes this application field in terms of available data and what sources of noise and domain shift is captured in the data. An important point to consider in

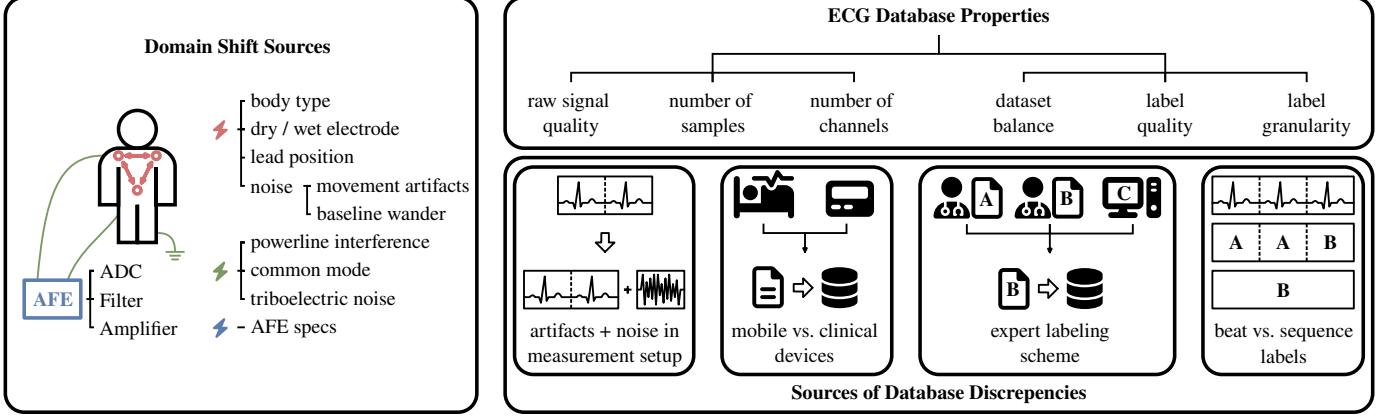


Fig. 2: An overview on the sources of domain shift in the application of ECG monitoring and the challenges of ECG data acquisition in general. Many databases differ in both measurement setup and labeling schemes making it challenging to combine them for DG.

the complexity of DG is that domain shift is not only caused by inter-patient differences [21], but also the experimental setup, which ranges from variations in lead positioning to contact impedance of electrodes resulting in deviating system responses for the captured data. The analog front end (AFE) further influences signal quality and introduces domain shift, when AFE configurations, i.e. ADC resolution etc., differ.

In principle, these perturbations can be modeled in a large enough corpus of datasets, which incorporates all varieties with sufficient sample sizes. However, biomedical data suffer from small number of labeled data [22] and unbalanced distribution of classes [23]. This is partially resulting from the scarcity of anomalies and the complexity of defining features, which even trained experts struggle to identify reliably [24]. The annotation of large amounts of data require large resources in terms of trained experts and time, which potentially affects the quality of provided labels [25]. Discrepancies in the labeling paradigm, e.g. beat vs. sequence labels or hybrid expert and machine-based labeling further complicate the combination of multiple datasets for training. Therefore, DG algorithm development mainly focus on homogeneous domain shift within one dataset, i.e. inter-patient paradigm [21], or multiple datasets with similar quality, i.e. (near) clinical measurement setup [26].

Figure 3 shows a conceptual overview of DG algorithms applied in ECG classification. One method is to select relevant data samples and add them into an active dataset, which is used to fine-tune or re-train the DNN [27], [28]. Since the new data incorporate samples in the target domain, the DNN is continuously adjusted towards new data. Another method adjusts the training process, such that features from different domains are aligned to each other, while features from different classes are separated [18]–[20], [29]–[32]. The main idea is that feature distributions from SD and TD are similar and should be robust against domain shift. The alignment process creates domain invariant features by minimizing distance metrics of those feature distributions. A third concept employs ensembles of classifiers, which classify each domain independently from each other through domain-specific branches in addition to

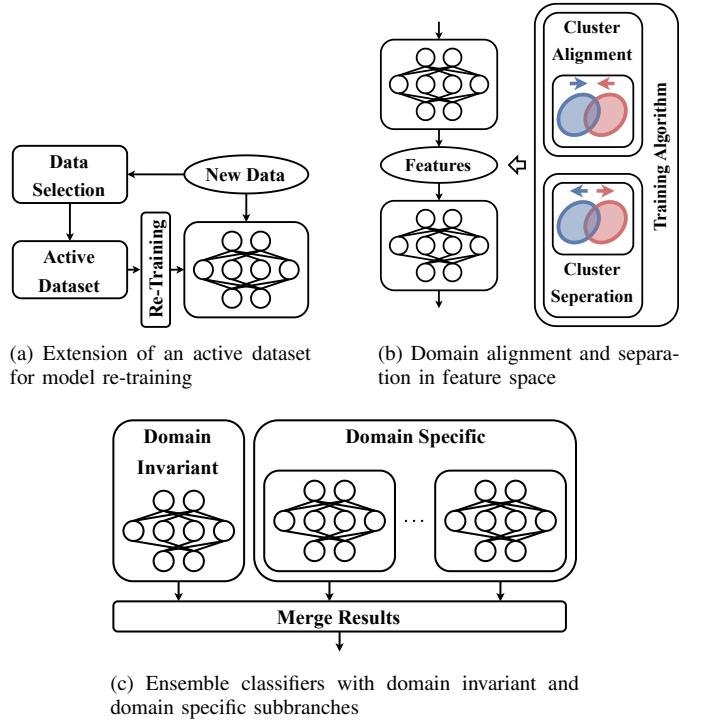


Fig. 3: Sketch of state-of-the-art DG algorithms for ECG classification

a general domain-invariant branch [17].

C. ECG accelerators

Classical ECG accelerators mainly focused on the efficient execution of classification models. These models range from traditional thresholding of ECG features, such as RR intervals [33], to recently exotic classifiers, such as spiking neural networks [34]–[37]. However, the majority of works investigates DNN variants such as convolutional neural networks (CNNs) [38]–[41], multi-layer perceptrons (MLPs) [42]–[44], long-short term memories (LSTMs) [45] and gated recurrent

units (GRUs) [46]. While all works aim to achieve ultra-low power consumption or energy per solution, the adopted hardware platform, i.e. ARM processor [45], FPGA [41] and ASICs [38]–[40], [42], [43], [46], varies. The complexity of the algorithm and their system architecture play a critical role in the choice. While processor architectures and FPGAs offer high flexibility in terms of programmability, the power consumption is generally several orders of magnitude higher than ASIC implementations ($>\mu\text{W}$ [41], [45] vs. nW range [40], [46]). Nevertheless, ASIC inference engines offer certain reconfigurability for model weights and architecture [38], [40], [46]. However, the architecture reconfiguration is usually limited to layer depth and width of convolution-based DNN layers. Generally, other processing components, such feature extraction, are fixed by the implemented architecture.

III. CORRECTION LAYER APPROACH

In this section, we detail the concept of CL as one additionally inserted layer in a pre-trained DNN. Here, the free trainable parameters of the CL are used to perform DG, while the pre-trained DNN remains frozen. As the training is limited to the CL only, we expect significant CC reduction compared to fine-tuning the whole DNN.

A. Motivation

Considering previous DG works (as summarized in Section II-B), it is evident that computational resources are not the primary optimization objective, since the extension of training dataset with new samples and domain specific DNN branches require scalable memory and computation resources. Furthermore, fine-tuning a DNN using domain alignment techniques on the edge either require on-device learning [47] or a distributed training setup to update all DNN parameters. Within this work, we investigate DG from the perspective of constrained complexity. A critical question is how much complexity is necessary to achieve generalization across domains.

To answer this question, we performed an experiment to evaluate the necessary complexity to achieve generalization for homogeneous and heterogeneous domain shift. In specific, we focused on the binary classification of atrial fibrillation (AF) and other (mainly normal) signals using a fully convolutional DNN. Training and validation is performed on the MIT-BIH Atrial Fibrillation Database (AFDB) [48] using 5-fold crossvalidation, while the Computing in Cardiology Challenge 2017 (CinC'17) [25] is added in the validation set. Here, inter-patient deviations in the dataset samples is used to incorporate homogeneous domain shift, where each fold contains a disjunct set of patients to guarantee unseen patients in the validation set. Due to the different data acquisition setup of the CinC'17 benchmark, those samples represent the heterogeneous domain shift compared to the AFDB reference. Figure 5 shows the employed network architecture and the feature distribution for the different datasets and the different DG techniques. Specifically, we chose patient-specific instance normalization (IN) as a regularization method with minimal additional parameters and contrastive learning (CTL) for feature alignment/separation using contrastive losses [49].

TABLE II: Validation F1 score of DG techniques based on homogeneous (AFDB) and heterogeneous (CinC'17) domain shift using 5-fold cross-validation.

	AFDB				CinC'17	
	Test - Known		Test - Unknown		Test - Unknown	
	N	AF	N	AF	N	AF
Baseline (w/o DG)	0.93 ± 0.03	0.90 ± 0.03	0.91 ± 0.07	0.88 ± 0.07	0.67 ± 0.01	0.17 ± 0.21
Instance Norm.	0.96 ± 0.02	0.93 ± 0.03	0.92 ± 0.10	0.91 ± 0.09	0.79 ± 0.02	0.78 ± 0.03
Contrastive Learning	0.97 ± 0.02	0.94 ± 0.02	0.96 ± 0.04	0.94 ± 0.05	0.80 ± 0.01	0.77 ± 0.02

Table II shows the detailed quality of service (QoS) of the classification. It is evident, that the baseline without DG is less accurate on unseen samples and performs poorly on CinC'17. IN and CTL improved QoS significantly in CinC'17, especially for AF. The feature distribution also indicates that a clear clustering of features is achieved through both methods. However, the QoS difference between IN and CTL is small, which motivates an emerging class of algorithms for HW co-optimization to fill the gap between high quality DG and no model finetuning. Inspired by the simplicity of IN as a linear transformation in each layer, we introduce correction layer (CL) as a single layer alternative in Section III-B. The key principle is to limit the modification to the addition of a single CL, such that the original DNN architecture and parameters remain untouched. This enables the reduction of CC and memory for training, which is reduced from the whole network to a single layer.

B. CL Algorithm

The general idea of CL is to insert a single layer into an existing DNN to transform the intermediate activations, such that the features after transformation are robust against domain shift. The layers themselves contain trainable parameters, which can be adjusted both in a supervised and unsupervised manner. Note that the resulting robust features can follow domain-specific models, such as language models in automatic speech recognition [50], but some applications, e.g. ECG classification, such domain-invariant expert features are not as clearly defined. From the structural point of view, the layer is not limited to typical DNN layers, such as convolution or fully-connected layers, but can take any form for transformation, which is trainable by e.g. backpropagation.

In our case, we opted for linear transformations due to their simplicity. One interesting property of linear transformations in this use case is, that they can be merged with adjacent linear transformations without any difference in the overall result. Since DNNs commonly consist of convolution and fully-connected layers, the integration of the transformation into existing weights is possible. In terms of computational complexity, the DNN inference generalized with merged CL is identical to its non-generalized baseline.

In our study, we implemented two types of CL: linear channel-wise and the linear inter-channel transform.

$$\mathbf{f}_{cw}(\mathbf{x}) = (\mathbf{w} + \mathbf{1}) \odot \mathbf{x} \quad (1)$$

$$\mathbf{f}_{ic}(\mathbf{x}) = (\mathbf{W} + \mathbb{I}) \cdot \mathbf{x} \quad (2)$$

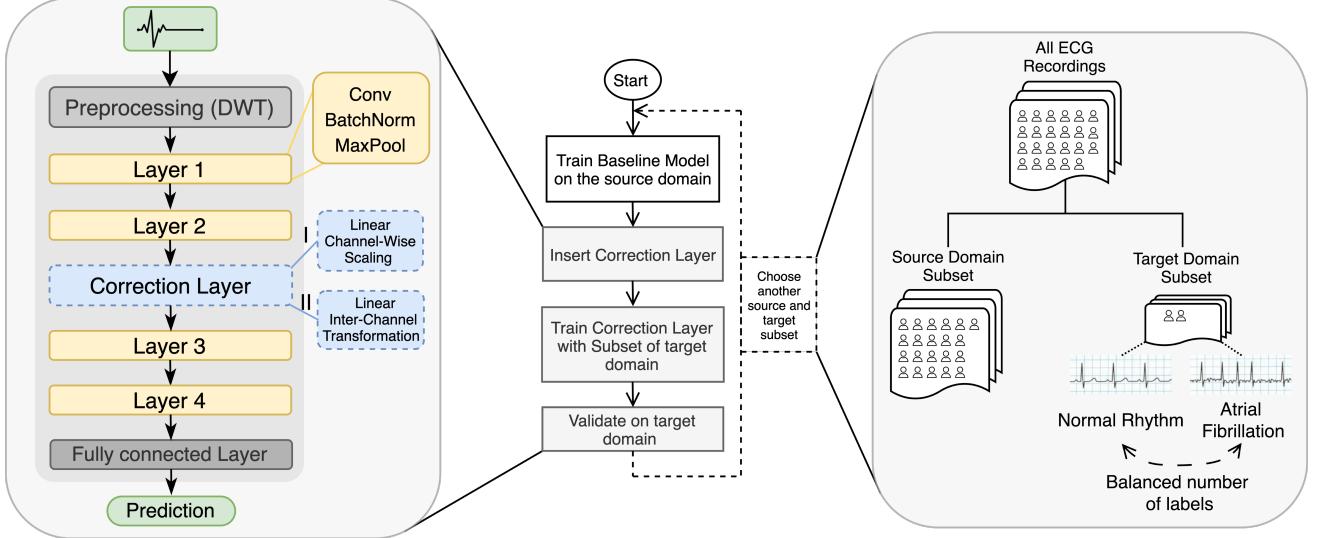


Fig. 4: Insertion of Correction Layer into CNN from [40]. The training and validation of DG capability is performed in two stages. The first stage trains the CNN without CL on SD. The second stage trains the CL only on a subset of TD, while the remainder of TD is used for validation. Both stages are performed for different combinations of patients in SD and are cross-validated for statistically robust QoS evaluation.

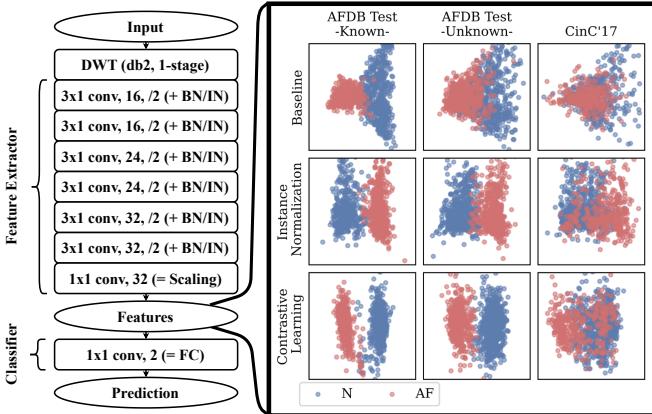


Fig. 5: Feature distribution of a fully convolutional DNN for IN and CTL. Features are mapped to the 2D plane using principle component analysis (PCA).

As noted in Eq. (1) and (2), CW scales each channel separately and IC additionally includes inter-channel dependencies. Here, IC offers more degrees of freedom to correct joint distribution discrepancies than CW, but also uses more trainable parameters.

C. CL QoS Robustness

To evaluate proposed CL, we chose the CNN architecture from [40], as it can solve the full CINC'17 benchmark with state-of-the-art QoS. Then, we trained the architecture for the binary classification use case with the AFDB data. Here, the data is split into SD and TD, such that a pre-trained model from the SD can be evaluated and then fine-tuned on the TD. The subsets are chosen by combining data of each patient,

such that the TD contains balanced number of labels¹, while the SD consists of the remaining patients. The balanced TD is necessary to ensure good fine-tuning performance on small sample sizes. The experiment is repeated for all permutations of one or two patients that guarantee balanced TD to generate statistically robust results. The evaluation within the TD subset is performed using stratified 5-fold cross-validation to ensure input data independent CL training performance. Note that in this experiment we specifically focus on inter-patient domain shift without loss in generality.

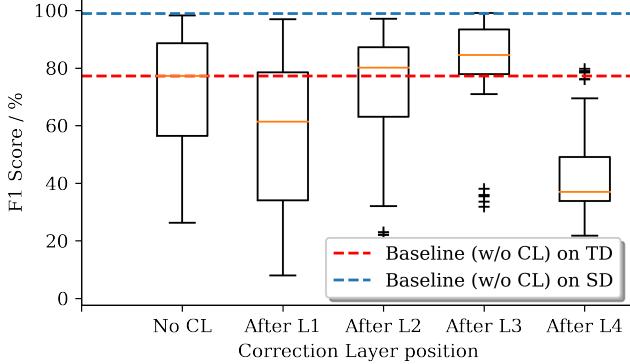
An important item to consider is that our experiment assumes the availability of TD data for training to explore the maximum capacity of CL in the context of DG. The modification of CL training towards unsupervised methods is expected to yield reduced QoS, but can be performed using metrics, which quantify the quality of generalized features and can be used as the loss function during training [20]. However, this is not investigated within the scope of this work.

D. Results and Discussion

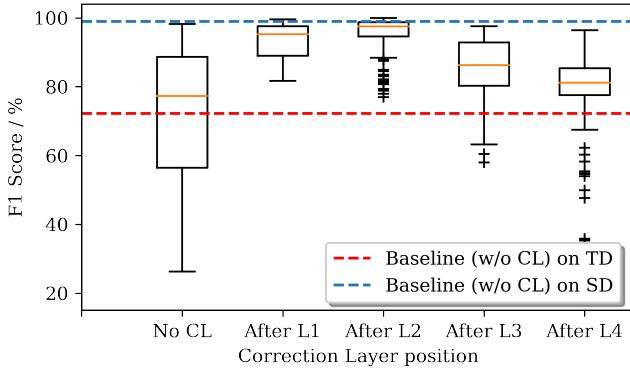
The performance on SD, TD before as well as TD after CL training is illustrated in Fig. 6 for all positions of the CL in between existing DNN layers. It is evident that all models perform very well on the SD (average: blue line) but the performance drops significantly on the unseen data from the TD (average: red line). Looking at the channel-wise CL, the improvement on QoS, i.e. F1-score, is selective to specific positions of the CL with some even underperforming the non-modified baseline. The inter-channel CL, however, shows consistent improvements in the F1-score, especially, when inserted in the center position of the baseline CNN with $\Delta F1 = 21.16\%$ compared to the average. Even though the

¹The number of labels from each class differs by max. 5 %

choice of used dataset is critical for a general interpretation (as discussed in Section II-B), the trend is robust for all combinations across train and test set patients in the investigated setup.



(a) Insertion of a single channel-wise CL after each layer



(b) Insertion of a single inter-channel CL after each layer

Fig. 6: Evaluation of CL performance regarding the type and position using CNN architecture from [40]. The blue and red line indicate the average performance on the SD and TD w/o CL training.

We also observe, that a larger degree of freedom, i.e. more trainable parameters in the CL, yield better results, since inter-channel CL clearly outperforms channel-wise CL. The good performance in the center position of Fig. 6b reflects previous observations about two opposing trends in feature distributions for DNN [6]. On the one hand, the task specificity of intermediate activations increase for layers close to the output, thus, decreasing the generalizing capability of those features. On the other hand, the discrepancies between input data decrease the deeper the intermediate features are relative to the input. In principle, the maximum capacity for generalization is, consequently, in the central layers of the DNN.

Although the results in Fig. 6 are based on CL training on the whole TD, we investigated whether the training process can be further simplified. To observe the necessary training samples for generalization, we repeated the experiment with less training samples. Specifically, we reduced the number of training samples per recording without any issues in training

convergence, whilst still keeping the same experiment setup as described in Section III-C. In the first try, a $2.99 \times$ reduction of training samples can be achieved with only 1% F1 score reduction. In the extreme case, even a $120.48 \times$ smaller training dataset is possible while tolerating an acceptable reduction of 6%. This result shows, that CL training does not require many training samples to yield acceptable QoS performance. Hence, test-time training of CL is a competitive option compared against state-of-the-art ECG DG methods utilizing active datasets, e.g. [27], [28], whilst still changing only CL parameter. Another big feature is that the chosen transformation is linear and can be merged with convolution and fully-connected layers (further evaluated in Section IV-A). It means that training is simply reduced to a single layer only, while it is observed in earlier experiment that this single layer fine-tuning requires only few samples. Consequently, the proposed CL showcases that constrained DNN training both in terms of model parameters and input samples yield state-of-the-art DG performance in the addressed usecase. In the end, we have successfully proven that the emerging class of HW-optimized DG algorithms has the potential to address the problem of DG with nearly negligible classification performance degradation compared to conventional approaches.

IV. HARDWARE COMPLEXITY

In a next step, we validate the predicted hardware efficiency of CL. As a reference for comparison, the status-quo of both DNN inference and training is considered. In the former, the integration of CL is investigated in an ultra-low power ECG DNN inference engine [40]. In the latter, on-device training is evaluated using high-level estimations for CC and memory overhead.

A. Integration of Correction Layer in an ECG accelerator

In the following, we implemented two logic designs for comparison: A reference design without CL (Ref) and a design with CL as a dedicated layer. The designs are synthesized in a 22 nm CMOS technology using commercial EDA tools, i.e. Synopsys Design Compiler. Both post-synthesis netlists are simulated at 0.8 V (nominal) and typical process corners at 2 MHz. To estimate the energy required to process each input sample the traces are recorded in the active calculation period. In our experiment, we chose the period of highest computational load, i.e. input sample triggering computation throughout the entire DNN until prediction sample is generated. Note that this worst case scenario deviates from the typical case as evaluated in [40], since it does not consider the idle periods in between input samples. This test configuration was selected to derive relevant quantitative comparison results while requiring reasonable simulation time and complexity.

The baseline HW architecture consists of activation memory, weight memory, a vector PE and control/multiplexing logic for activation routing. In the reference architecture, the activation memory and weight memory is split into shift registers and SRAM respectively, due to stream processing architecture and the utilized output stationary dataflow to reuse the vector PE element.

The integration of an additional layer, therefore requires simply a larger SRAM for the additional weights and another set of shift registers for the input activations of the CL. A key difference is the mapping of the matrix-vector multiplication (from Eq. (2)) to the existing vector PE. This is achieved through loop tiling and corresponding control logic modifications. For instance, the computation of one row of a 24×24 matrix with the input vector of 24 elements is treated the same as a 5×1 convolution with 5 input channels. The number of output elements, i.e. 24 rows of the weight matrix, correspond to the number of output channels in the convolution and the output samples are fed into the shift registers of the next layer.

Similar to the batch normalization layers, adjacent linear operations, such as the matrix-vector multiplication of the CL and the convolution of a convolution layer, can be merged into a single operation, while still computing the arithmetically identical output activations. Conceptually, this results in a third design, i.e. the CL merged into the adjacent convolution layer.

TABLE III: Post-synthesis results of reference ECG accelerator with and without integrated CL

	Train Acc. (%)	Test Acc. (%)	Area (μm^2)	Seq. Cells	Comb. Cells	Max. Freq. (MHz)	Energy (nJ)
CL Design	98.98	97.16	19.9k	19.7k	6.9k	9.96	68.21
CL Design (integr.)			19.3k	19.4k	6.6k	9.97	65.14
Reference Design	99.01	95.15					

Table III shows the post-synthesis results of the implemented logic designs. It is evident, that the CL design performs more robust across both train and test set generalizing across patient groups. If the CL is implemented explicitly, more sequential cells are synthesized to store CL input activations and more combinational logic is necessary for control/routing logic. Therefore, the overall area and energy is increased slightly by 3.1% and 4.7%, respectively. In the case that CL is merged into the convolution, the reference design can be reused with different weights in one layer. In terms of the design's HW key performance indicators (KPIs), they are identical to the reference design with the classification performance of the CL design. Hence, no compromises need to be made for CL integration in the DNN inference.

B. Backpropagation On-Chip

Considering inference alone does not reveal the strength of proposed CL. Therefore, in a next step, we estimate expected CC and memory requirements of CL in comparison to conventional fine-tuning during deployment.

As detailed in [47], the weight updates in the training step are calculated based on the input activations $x_i^{(l)}$, weights $w_{ij}^{(l)}$ and the gradients of each layer l , where i, j indicate the neuron in the input and output of the layer, respectively. In specific, the weight increment $\Delta w_{ij}^{(l)}$ is calculated using one variant of the stochastic gradient descent algorithm, which depend on $x_i^{(l)}$ and $w_{ij}^{(l)}$ through the gradient of the loss function $\partial L / \partial w_{ij}^{(l)}$ (see Eq. (3) and (4)).

$$w_{ij}^{(l)} \rightarrow w_{ij}^{(l)} + \Delta w_{ij}^{(l)} \quad (3)$$

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial x_j^{(l)}} \frac{\partial x_j^{(l)}}{\partial w_{ij}^{(l)}} \quad (4)$$

Figure 7 visualizes the memory and CC reductions of CL training compared to conventional DNN fine-tuning reference in an example. For the reference, all partial derivatives need to be calculated in all layers and, therefore, all activations need to be stored as well for the weight update. For CL training, only the weights and activations of the CL need to be stored in addition to the reference DNN, while the partial derivatives $\partial L / \partial x_j^{(l)}$ required for calculation do not need permanent buffers. They are calculated recursively starting from the loss at the output layer, while the activations of other layers are not required for their computation. We approximate CC with multiply-and-accumulate (MAC) operations similar to complexity estimations in the forward pass [51]. We further consider different layer types, such as fully connected (FC) layers and convolution (CONV) layers with and without subsequent pooling, to increase the granularity of our estimation model.

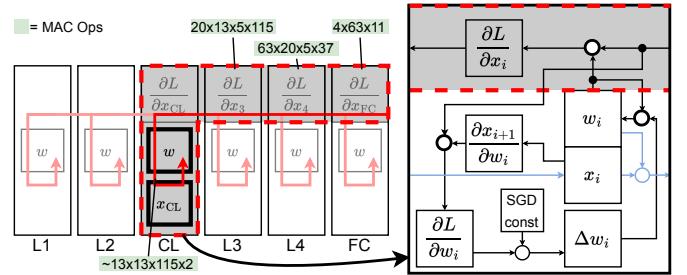
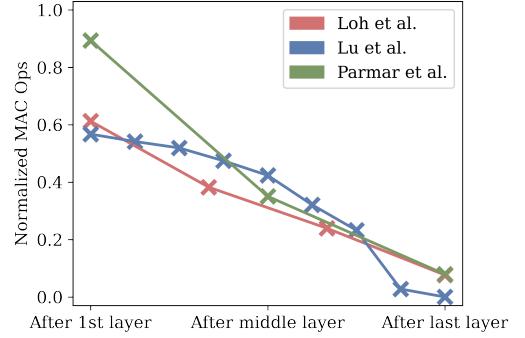


Fig. 7: Simplified concept of CL CC and memory requirements in the example CNN from [40]. Additional memory is marked in bold solid lines and the CC for training is indicated in the area marked with red dashed lines. The MAC operations for each subcomponent is highlighted in green.

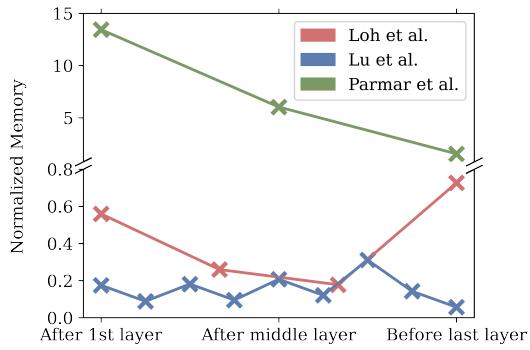
Figure 8 shows results of CC and memory estimations for different positions of the CL within a variety of DNN architectures, which are suitable for HW acceleration on embedded platforms. In this comparison, inter-channel CL are used as an example, since they showed the best QoS performance in the evaluation in Section III.

In general, fine-tuning the complete model requires most MAC operations compared to the insertion of a CL regardless of the DNN architecture (see Fig. 8a). Further it is visible, that the number of MAC operations decrease the closer the CL is inserted to the final output layer. This trend is a consequence from the recursive calculation of $\partial L / \partial x_j^{(k)}$. The overall CC, however, is dependent on the DNN model architecture, as CL size is dependent on the dimensions of its input.

The additional memory required for CL, however, does not show such consistent trend (see Fig. 8b). While CL in the CNN architectures from [52] and [40] show consistently less memory overhead than the DNN fine-tuning, the MLP architecture of [44] includes a higher memory requirement



(a) Normalized MAC operations for training over the position of inserted CL in the DNN (left: input, right: output)



(b) Normalized memory overhead for training over the position of inserted CL in the DNN (left: input, right: output)

Fig. 8: Memory and CC of backpropagation for HW accelerators Loh et al. [40], Lu et al. [52] and Parmar et al. [44]. The complexity is normalized against the reference case, in which the whole DNN is fine-tuned.

for CL training. Here, the number of parameters necessary for the CL, i.e. square of intermediate activations of that layer, exceed the storage saved by omitting the buffer for the frozen layers. As the relative difference of CL parameters decrease for deeper networks, this trend is an exception for shallow DNNs. Furthermore, the memory overhead is mainly dependent on the size of inserted CL, i.e. its input dimensions. For instance, the intermediate activations in the CNN model of [40] consistently decrease the closer the CL is inserted to the output layer. However, the memory increases when inserted directly before the final FC layer, as its input dimensions, i.e. feature dimension and input channels, are much larger than in the previous layers. A different example is the model of [52], where the input dimensions of all layers are quite consistent. Here, the aforementioned memory reductions with max pooling layers are visible in nearly every second layer.

All in all, a trade-off needs to be found between achieved QoS, CC and memory. Considering the best design point from the QoS evaluation (see Section III), i.e. CL after L2, a CC reduction of $2.5 \times$ and a memory reduction of more than $3 \times$ is observed. Hence, CL training provide a more efficient alternative to fine-tuning, while achieving competitive classification performance on the target domain.

V. CONCLUSION

Practical deployment of DNN models in mobile devices require robust methods to cope with the domain shift problem. The main challenge is the combination of the two independent disciplines of DNN HW acceleration and DG approaches on the algorithm level. Especially, sensitive applications such as ECG classifications for health monitoring benefit from algorithm-hardware co-designed systems for DG on-the-edge, as domain shift is inherent in practical scenarios and privacy concerns limit the data transmission to central servers. To address this challenge, this work presented DG methods specifically for the deployment on HW accelerators using ECG processing as the application context. The variety of DG methods require modifications on the pre-trained DNN, while the majority aims to fine-tune the pre-trained DNN to generalize across domains. In the case of ECG classification, the scarcity of data and high quality labels limit the capability of DG methods to fully capture all variants of domain shift, although state-of-the-art methods are capable of solving inter-patient domain shift effectively. As a first step towards co-optimized DG methods, we introduced ‘‘correction layers’’ (CLs) as a low complexity DG method to solve inter-patient domain shift. This is achieved by freezing a pre-trained DNN, while adding a single trainable CL for feature normalization. Our evaluation shows that CL with inter-channel transformations provide robust QoS improvements of $\Delta F1 \geq 20\%$. These improvements in classification performance require no hardware overhead during inference and are complemented by CC and memory reductions of more than $2.5 \times$ and $3 \times$ during training, respectively. In the end, the CL study shows that the co-optimization of algorithm and hardware yield state-of-the-art ECG classification results on inter-patient DG with minimal modification on existing ECG accelerators.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘Imagenet classification with deep convolutional neural networks,’’ *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [2] M. Zeineldin, J. Xu, C. Luscher *et al.*, ‘‘Conformer-based hybrid ASR system for switchboard dataset,’’ in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022.
- [3] C. Latotzke and T. Gemmeke, ‘‘Efficiency versus accuracy: A review of design techniques for dnn hardware accelerators,’’ *IEEE Access*, vol. 9, pp. 9785–9799, 2021.
- [4] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer *et al.*, Eds., *Dataset Shift in Machine Learning*, ser. Neural Information Processing series. London, England: MIT Press, Jun. 2022.
- [5] K. Zhou, Z. Liu, Y. Qiao *et al.*, ‘‘Domain generalization: A survey,’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.
- [6] J. Yosinski, J. Clune, Y. Bengio *et al.*, ‘‘How transferable are features in deep neural networks?’’ *Advances in neural information processing systems*, vol. 27, 2014.
- [7] S. Zhao, X. Yue, S. Zhang *et al.*, ‘‘A review of single-source deep unsupervised visual domain adaptation,’’ *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 473–493, Feb. 2022.
- [8] F. Pourpanah, M. Abdar, Y. Luo *et al.*, ‘‘A review of generalized zero-shot learning methods,’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.
- [9] F. Zhuang, Z. Qi, K. Duan *et al.*, ‘‘A comprehensive survey on transfer learning,’’ *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan 2021.
- [10] H.-C. Seo, S. Oh, H. Kim *et al.*, ‘‘ECG data dependency for atrial fibrillation detection based on residual networks,’’ *Scientific Reports*, vol. 11, no. 1, Sep. 2021.

- [11] M. Long, Y. Cao, J. Wang *et al.*, “Learning transferable features with deep adaptation networks,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML’15. JMLR.org, 2015, p. 97–105.
- [12] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML’13. JMLR.org, 2013, p. I–10–I–18.
- [13] Y. Sun, X. Wang, Z. Liu *et al.*, “Test-time training with self-supervision for generalization under distribution shifts,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 9229–9248.
- [14] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, jul 2019.
- [15] R. Krishnan, P. Rajpurkar, and E. J. Topol, “Self-supervised learning in medicine and healthcare,” *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1346–1352, Aug. 2022.
- [16] D. Li, Y. Yang, Y.-Z. Song *et al.*, “Learning to generalize: Meta-learning for domain generalization,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.
- [17] F. Deng, S. Tu, and L. Xu, “Multi-source unsupervised domain adaptation for ECG classification,” in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Dec. 2021.
- [18] Y. Bazi, N. Alajlan, H. AlHichri *et al.*, “Domain adaptation methods for ECG classification,” in *2013 International Conference on Computer Medical Applications (ICMA)*. IEEE, Jan. 2013.
- [19] M. Chen, G. Wang, Z. Ding *et al.*, “Unsupervised domain adaptation for ECG arrhythmia classification,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, Jul. 2020.
- [20] G. Wang, M. Chen, Z. Ding *et al.*, “Inter-patient ECG arrhythmia heartbeat classification based on unsupervised domain adaptation,” *Neurocomputing*, vol. 454, pp. 339–349, Sep. 2021.
- [21] M. Sraithi, Y. Jabrane, and A. Atlas, “An overview on intra- and inter-patient paradigm for ECG heartbeat arrhythmia classification,” in *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*. IEEE, Jun. 2021.
- [22] T. Mehari and N. Strothoff, “Self-supervised representation learning from 12-lead ECG data,” *Computers in Biology and Medicine*, vol. 141, p. 105114, Feb. 2022.
- [23] S. Wang, Y. Dai, J. Shen *et al.*, “Research on expansion and classification of imbalanced data based on smote algorithm,” *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.
- [24] P. Kirchhof, S. Benussi, D. Kotecha *et al.*, “2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS,” *European Heart Journal*, vol. 37, no. 38, pp. 2893–2962, Aug. 2016.
- [25] G. D. Clifford, C. Liu, B. Moody *et al.*, “Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017,” in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.
- [26] E. A. P. Alday, A. Gu, A. J. Shah *et al.*, “Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020,” *Physiological measurement*, vol. 41, no. 12, p. 124003, 2020.
- [27] M. A. Rahhal, Y. Bazi, H. AlHichri *et al.*, “Deep learning approach for active classification of electrocardiogram signals,” *Information Sciences*, vol. 345, pp. 340–354, Jun. 2016.
- [28] G. Wang, C. Zhang, Y. Liu *et al.*, “A global and updatable ECG beat classification system based on recurrent neural networks and active learning,” *Information Sciences*, vol. 501, pp. 523–542, Oct. 2019.
- [29] Y. Ganin, E. Ustinova, H. Ajakan *et al.*, “Domain-adversarial training of neural networks,” *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [30] W. Yin, X. Yang, L. Li *et al.*, “Self-adjustable domain adaptation in personalized ECG monitoring integrated with IR-UWB radar,” *Biomedical Signal Processing and Control*, vol. 47, pp. 75–87, jan 2019.
- [31] H. Hasani, A. Bitarafan, and M. Soleymani, “Classification of 12-lead ECG signals with adversarial multi-source domain generalization,” in *2020 Computing in Cardiology Conference (CinC)*. Computing in Cardiology, Dec. 2020.
- [32] Z. Shang, Z. Zhao, H. Fang *et al.*, “Deep discriminative domain generalization with adversarial feature learning for classifying ECG signals,” in *2021 Computing in Cardiology (CinC)*. IEEE, Sep. 2021.
- [33] Y.-P. Chen, D. Jeon, Y. Lee *et al.*, “An injectable 64 nW ECG mixed-signal SoC in 65 nm for arrhythmia monitoring,” *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 375–390, Jan. 2015.
- [34] F. C. Bauer, D. R. Muir, and G. Indiveri, “Real-time ultra-low power ECG anomaly detection using an event-driven neuromorphic processor,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1575–1582, Dec. 2019.
- [35] H. Chu, Y. Yan, L. Gan *et al.*, “A neuromorphic processing system with spike-driven SNN processor for wearable ECG classification,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 4, pp. 511–523, Aug. 2022.
- [36] Y. Liu, Z. Wang, W. He *et al.*, “An 82nw 0.53pj/SOP clock-free spiking neural network with 40 μ s latency for AloT wake-up functions using ultimate-event-driven bionic architecture and computing-in-memory technique,” in *2022 IEEE International Solid- State Circuits Conference (ISSCC)*. IEEE, Feb. 2022.
- [37] R. Mao, S. Li, Z. Zhang *et al.*, “An ultra-energy-efficient and high accuracy ECG classification processor with SNN inference assisted by on-chip ANN learning,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 5, pp. 832–841, Oct. 2022.
- [38] J. Liu, Z. Zhu, Y. Zhou *et al.*, “4.5 BioAIP: A reconfigurable biomedical AI processor with adaptive learning for versatile intelligent health monitoring,” in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*. IEEE, Feb. 2021.
- [39] Z. Wang, Y. Liu, P. Zhou *et al.*, “A 148-nW reconfigurable event-driven intelligent wake-up system for AloT nodes using an asynchronous pulse-based feature extractor and a convolutional neural network,” *IEEE Journal of Solid-State Circuits*, vol. 56, no. 11, pp. 3274–3288, Nov. 2021.
- [40] J. Loh and T. Gemmeke, “Dataflow optimizations in a sub-uW data-driven TCN accelerator for continuous ECG monitoring,” in *2022 IEEE Nordic Circuits and Systems Conference (NorCAS)*. IEEE, Oct. 2022.
- [41] J. Lu, D. Liu, X. Cheng *et al.*, “An efficient unstructured sparse convolutional neural network accelerator for wearable ECG classification device,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 11, pp. 4572–4582, Nov. 2022.
- [42] S. Yin, M. Kim, D. Kadetotad *et al.*, “A 1.06-uw smart ECG processor in 65-nm CMOS for real-time biometric authentication and personal cardiac monitoring,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 8, pp. 2316–2326, Aug. 2019.
- [43] Y. Zhao, Z. Shang, and Y. Lian, “A 13.34 uw event-driven patient-specific ANN cardiac arrhythmia classifier for wearable ECG sensors,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 2, pp. 186–197, Apr. 2020.
- [44] R. Parmar, M. Janveja, J. Pidanic *et al.*, “Design of DNN-based low-power VLSI architecture to classify atrial fibrillation for wearable devices,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 3, pp. 320–330, Mar. 2023.
- [45] G. Sivapalan, K. K. Nundy, S. Dev *et al.*, “ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 1, pp. 24–35, Feb. 2022.
- [46] M. Jobst, J. Partzsch, C. Liu *et al.*, “ZEN: A flexible energy-efficient hardware classifier exploiting temporal sparsity in ECG data,” in *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, Jun. 2022.
- [47] J. Lee and H.-J. Yoo, “An overview of energy-efficient hardware accelerators for on-device deep-neural-network training,” *IEEE Open Journal of the Solid-State Circuits Society*, vol. 1, pp. 115–128, 2021.
- [48] G. Moody, “A new method for detecting atrial fibrillation using rr intervals,” *Proc. Comput. Cardiol.*, vol. 10, pp. 227–230, 1983.
- [49] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 2. IEEE, 2006, pp. 1735–1742.
- [50] J. Li *et al.*, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [51] A. Parashar, P. Raina, Y. S. Shao *et al.*, “Timeloop: A systematic approach to dnn accelerator evaluation,” in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2019, pp. 304–315.
- [52] J. Lu, D. Liu, Z. Liu *et al.*, “Efficient hardware architecture of convolutional neural network for ECG classification in wearable healthcare device,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 7, pp. 2976–2985, Jul. 2021.