UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Machine Learning Techniques for Low-Power Mobile Health Systems

Permalink

https://escholarship.org/uc/item/9wx191sv

Author

Chen, Luke

Publication Date

2023

Copyright Information

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Machine Learning Techniques for Low-Power Mobile Health Systems

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Electrical and Computer Engineering

by

Luke Chen

Thesis Committee:
Professor Mohammad Al Faruque, Chair
Assistant Professor Peter Tseng
Assistant Professor Salma Elmalaki

2023

DEDICATION

To my family for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

ABSTRACT OF THE THESIS

Machine Learning Techniques for Low-Power Mobile Health Systems

By

Luke Chen

Master of Science in Electrical and Computer Engineering

University of California, Irvine, 2023

Professor Mohammad Al Faruque, Chair

With ever-growing interests in personalized physical and mental healthcare, especially with the recent COVID-19 pandemic, along with the proliferation of applied machine learning (ML), automated health monitoring and prognosis using wearable devices with ML algorithms are increasingly more relevant. However, modern deep learning (DL) frameworks with state-of-the-art performances are unable to meet the memory and energy constraints of wearable devices. As such, there is a need for the design of efficient and effective ML models that can operate on these constrained devices while still achieving acceptable performance thresholds. This thesis presents two methodologies for effective and efficient ML under the wearable device setting, specifically a feature-augmented hybrid convolutional neural network architecture for stress monitoring using a wrist-based photoplethysmography sensor and a neural contextual-bandits-based dynamic sensor selection framework for cardiovascular disease detection with body-area-network of electrocardiogram sensor. Through extensive empirical studies, we find that our methods satisfy the constrained device settings while maintaining task performance and in some cases outperforming related ML and deep learning works. Additionally, we performed feasibility analysis on real embedded micro-controller hardware where run-time memory and energy profiling was measured and reported.

Chapter 1

Introduction

1.1 Background

1.1.1 Stress Recognition

The COVID pandemic has caused more than 6.62 Million global deaths since its beginning in 2019 up until the last quarter of 2022 according to Google statistics. Unsurprisingly, there have been increased levels of mental stress among people across the globe where the American Psychological Association (APA) warned of "A National Mental Health Crisis" in their October 2020 report [2]. A follow-up 2021 survey on the effects of the pandemic with 3013 adults reported that 61% experienced undesired weight changes, 67% sleeping problems, and 25% encountered mental health disorders [3]. Stress is a physiological state that hampers mental health and is associated with the fight-or-flight response due to surges in body chemical and hormonal levels. Stress can be **Physical** (exercising); **Cognitive** (problem-solving); and **Emotional** (fear, anxiety). According to the APA, there are 3 further categorizations of stress associated with its experienced frequency, specifically, **Acute**

or short-term everyday stress, **Episodic Acute** or repeating acute stress, and **Chronic** stress or long-term stress that last months or even years. Chronic stress is one of the major associations with adverse effects on physical health which include clinical depression, sleep deprivation/oversleeping, abnormal body weight changes, cardiovascular diseases, and even suicide. As such, recognition of emotional stress has become more crucial now than ever.

1.1.2 Cardiovascular Disease Detection

Amongst the adverse effects of stress mentioned previously, one of the most concerning is the association between stress and increased risk of cardiovascular diseases (CVD), especially for those with pre-existing conditions [47, 32, 46, 49]. CVDs are one of the leading causes of mortality worldwide where the age-adjusted death rate of CVD was 219.4 per 100,000 in 2017, or 859,125 dead and 2.2 million hospitalized in the U.S alone [5], which is more than cancer and chronic lung disease combined [4]. Among CVDs, there are different types of cardiovascular anomalies including myocardial infarction (MI), conduction disturbance (CD), ST/T-changes (STTC), and hypertrophy (HYP) to name a few, each with its own unique abnormal heart rhythm characteristic.

1.1.3 Advancements in Wearable Technologies

Recent advances in wearable technology and data analysis techniques have enabled the continuous and automated monitoring and detection of stress and cardiovascular diseases using various physiological sensors and machine learning or deep learning algorithms [34, 35]. Among the various devices, wearable wrist devices, like smartwatches, are most popular due to their minimal form factor for convenience and aesthetic appeal. Sensors common to both applications include electrocardiogram (ECG) and photoplethysmography (PPG) which both capture heart-related bio-metrics. Other sensors such as accelerometers (ACC), electroder-

mal activity (EDA), electromyograms (EMG), respiration (RESP), and temperature sensors (TEMP) have been used for stress detection. Popular medical-grade commercial devices such as the Empatica E4, and Respiban include most of these sensors in a compact form factor that allows data collection in a more natural setting whereas clinical-grade devices are bulky and often restrict subjects to a lab environment.

1.1.4 Advancements in Machine Learning Techniques

With the availability of large volumes of healthcare datasets and advancements in deep learning techniques, systems are now potentially well-equipped in diagnosing many health problems. Among these deep learning models, convolutional neural networks (CNN) are often used for their efficient and automatic extraction of spatial features and memory-based recurrent neural network (RNN) architectures with their variants, namely, the long short-term memory (LSTM) models and gated recurrent units (GRU), have been studied for extracting temporal patterns from sequential time-series healthcare data for various diagnostic categories and classifications. The effectiveness of ML techniques has been demonstrated through various healthcare applications [36, 37, 12, 9]. For example, CNNs in combination with LSTMs used for myocardial infarction detection have achieved high prediction accuracy with only raw ECG data [13, 33]. Similarly for multi-modal human activity recognition, the addition of attention mechanisms improved feature extraction by focusing on meaningful sensor inputs to achieve state-of-the-art (SOTA) performances [24, 10]. Success has also been found in sleep stage classification where desirable performances were achieved using only single lead EEG data [11, 18]. These varied successes demonstrate the effectiveness of deep learning algorithms in their ability to extract high-level abstract features and capture spatial-temporal correlations and are the key to achieving outstanding performance in healthcare applications and other fields.

Chapter 2

Feature Augmented Hybrid CNN for Stress Recognition Using Wrist-based Photoplethysmography Sensor

2.1 Photoplethysmography

As introduced in Section 1.1.3, PPG sensors are one of the primary sensors used in wearable devices to capture cardiovascular biometrics. It is a noninvasive optical-based sensor that captures differences in blood volume pulse (BVP), though more informative features can be extracted from it such as heart rate, oxygen saturation, blood pressure, etc. More recently, it has also been shown to be a strong biomarker for stress and emotion recognition [40]. Therefore, this chapter focuses on detecting stress using a wrist-based PPG sensor suitable for daily monitoring via consumer-grade wristwatches. Moreover, state-of-the-art works have used either classical machine learning algorithms to detect stress or emotion using the aforementioned hand-crafted features [22] or have used deep learning algorithms like CNNs

to automatically extract features. In this chapter, we introduce a novel hybrid CNN (H-CNN) that uses both hand-crafted features and automatically extracted features by CNN to detect stress. In this work, I analyzed the different performance effects of standard features compared to CNN features as well as analyzing the performance benefits of using a CNN compared to classical ML algorithms. We demonstrate the effectiveness of our hybrid approach using wrist-based BVP signal from the WESAD [40] dataset.

The novel contributions of our method are as follows:

- Propose a novel hybrid CNN (H-CNN) classifier for stress detection using a wrist-based PPG sensor. It uses both handcrafted features and automatically extracted features by CNN to detect stress.

- Validation of our proposed approach using BVP signal from WESAD [40] dataset collected through wrist-based PPG.

- Evaluation on the benchmark WESAD dataset shows that, for 3-class classification (Baseline vs. Stress vs. Amusement), our proposed H-CNN outperforms traditional classifiers and normal CNN by $\approx 5\%$ and $\approx 7\%$ accuracy, and $\approx 10\%$ and $\approx 7\%$ macro F1 score, respectively. Also for 2-class classification (Stress vs. Non-stress), our proposed H-CNN outperforms traditional classifiers and normal CNN by $\approx 3\%$ and $\approx 5\%$ accuracy, and $\approx 3\%$ and $\approx 7\%$ macro F1 score, respectively.
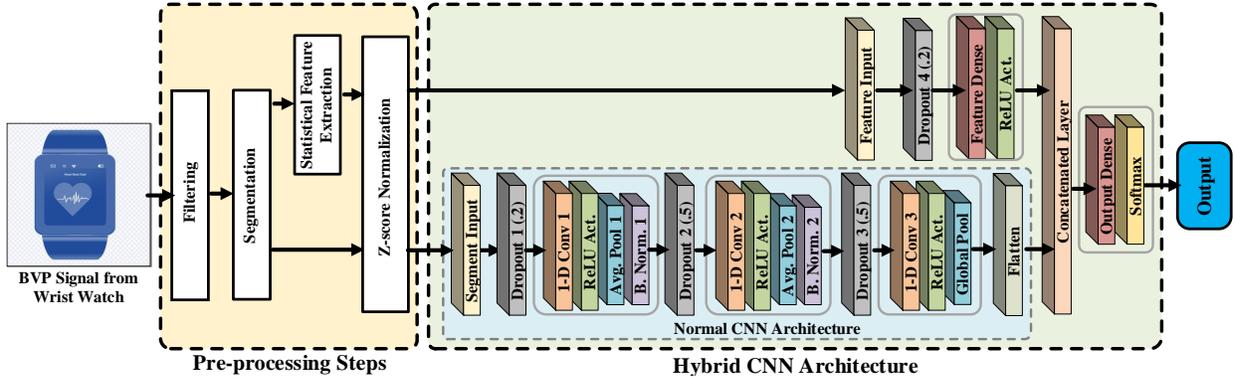
Figure 2.1: Overview of Our Proposed Methodology

2.2 Methodology

2.2.1 Pre-processing

As shown in Figure 2.1, the pre-processing steps start with filtering the raw BVP signal. We filter the raw BVP signal by a butter-worth bandpass filter of order 3 with cutoff frequencies (f_1=.7 Hz and f_2=3.7 Hz). We take into account the heart rate at rest ($\approx$40 BPM) or high heart rate due to exercise scenarios or tachycardia ($\approx$220 BPM) following the method mentioned in [39].

2.2.2 Segmentation

The filtered signal is segmented by a window of 60 seconds of data following the paper that introduced the WESAD dataset [40]. We use a sliding length of 5 seconds in between the segments. Each segment contains 3840 samples as the sampling rate of the BVP signal is 64 Hz.

Table 2.1: List of Extracted Features

Feature Symbol	Feature Names
μ_{HR}, σ_{HR}	Mean and Standard Devaiation of HR
μ_{HRV}, σ_{HRV}	Mean and Standard Devaiation of HRV
$NN50, pNN50$	Number and percentage of HRV intervals differing more than 50 ms
rms_{HRV}	Root mean square of the HRV
f_{HRV}^{x} $x \in ULF, LF, HF, UHF$	Energy in different frequency component of the HRV
$f_{HRV}^{LF/HF}$	Ratio of LF and HF component
$\sum_{x}^{f}$ $x \in ULF, LF, HF, UHF$	$\sum$ of the frequnecy components in ULF-HF
rel_{x}^{f}	Relative power of freq. components
LF_{norm}, HF_{norm}	Normalised LF and HF component

Heart Rate (HR), Heart Rate Variability (HRV)

2.2.3 Feature extraction

The first step of feature extraction is the detection of heartbeats. Once the peaks are detected, different time domain and frequency domain features are extracted based on the location of the peaks. We extract the time and frequency domain features as in [40] to ensure a fair comparison of our H-CNN classifier against the traditional machine learning classifiers used in the WESAD paper. We use the same frequency bands - ultra-low (ULF: 0.01-0.04 Hz), low (LF: 0.04-0.15 Hz), high (HF: 0.15-0.4 Hz) and ultra-high (UHF: 0.4-1.0 Hz) band as in [40] to calculate different frequency domain features. The list of extracted features is given in Table 2.1.

2.2.4 Hybrid CNN (H-CNN) Architecture

The normalized BVP segments and the corresponding features for each segment are passed to our H-CNN architecture following the framework in Figure 2.1.

The H-CNN architecture has two input layers - Segment and Feature input. The segment

input layer is followed by a dropout layer (with a 20% dropout rate) which is then followed by 3 convolution blocks. The first and second convolution blocks have - convolution, $ReLU$ activation, average pooling, and batch normalization layers. Both first and second convolution block is followed by dropout layers with 50% dropout rate which are added to reduce overfitting. The third convolution block has one convolution layer followed by a global average pooling layer which is also used to reduce the overfitting of the CNN. For the normal CNN architecture, the output of the global average pooling layer is directly fed to the output dense layer followed by a $Softmax$ activation. However, for the H-CNN architecture, the output of the global average pooling layer is concatenated with the feature-dense layer. Finally, the concatenated layer is fed to the output dense layer that is followed by the $Softmax$ activation. The details of our H-CNN architecture are shown in Table 2.2. As shown in Table 2.2, the total number of parameters required to classify a segment is $6846+(13*n_c)$, where n_c is the number of output classes. In this paper, we perform both 2-class (Stress vs. Non-stress) and 3-class (Baseline vs. Stress vs. Amusement) classification from the WESAD dataset.

2.3   Experimental Evaluation

2.3.1   Dataset

The WESAD dataset is used for the validation of our proposed methodology as it is the only publicly available dataset that contains wrist-based PPG sensor data for stress and affect detection. Although the dataset contains data for a total of 15 subjects from both chest (RespiBAN) and wrist (Empatica E4) worn sensors, we are only interested in using the wrist-based BVP signal collected through the PPG sensor. The dataset is labeled for 3 types of classes - baseline (neutral), amusement, stress.

Table 2.2: Hybrid CNN Architecture Details

Layer Name	Kernel Size	Stride Size	Act. Func.	Output Shape	# of Param.
Seg. Inp.	-	-	-	3840x1	0
D.O. 1	-	-	-	3840x1	0
Conv 1	64	4	ReLU	945x8	520
Pool 1	4	4	-	236x8	0
B.N. 1	-	-	-	236x8	32
D.O. 2	-	-	-	236x8	0
Conv 2	32	2	ReLU	103x16	4112
Pool 2	4	4	-	25x16	0
B.N. 2	-	-	-	25x16	64
D.O. 3	-	-	-	25x16	0
Conv 3	16	1	ReLU	10x8	2056
G. Pool	4	4	-	8	0
Flatten	-	-	-	8	0
Feat. Inp.	-	-	-	19	0
D.O. 4	-	-	-	19	0
Feat. Den.	-	-	ReLU	4	80
Concate	-	-	-	12	0
Out. Den.	-	-	SM	n_c	$13 \ast n_c$
Total Number of Parameters					**6846+(13*n_c)**

Segment Input (Seg. Inp.), Dropout (D.O.), Batch Normalization (B.N.), Global Average Pooling (G. Pool), Feature Input (Feat. Inp.), Feature Dense (Feat. Den.), Output Dense (Out. Den.), Softmax (S.M.)

2.3.2 Performance Metric

As the number of segments for different classes in the dataset is highly imbalanced, only classification accuracy is not appropriate to measure performance. Therefore, the F1 score provides a better measure that balances precision and recall performance. To ensure a fair comparison with our related work in [40], we use a macro F1 score where each class is given equal importance. The metrics used for evaluation are given below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$Macro\ F_1 = \frac{1}{n_c} \sum_{i}^{n_c} 2 * \frac{Precision_i.Recall_i}{Precision_i + Recall_i} \tag{2.4}$$

Where TP, TN, FP, FN represents True Positives, True Negatives, False Positives, and False Negatives respectively. The classes are indexed by i, and n_c is the number of output classes.

2.3.3 Model Training and Evaluation

We train our normal CNN and H-CNN classifiers with a batch size of 500. The models are trained for 200 epochs with an early stopping mechanism having a patience value of 70. We monitor the validation recall value to select the best model from the epochs. To ensure proper training for the imbalance dataset, we assign class weights to each class using the following formula in Eq. 2.5.

$$w_i = \frac{1}{N_i} * \frac{N}{n_c} \tag{2.5}$$

Here, w_i, and N_i represent the class weight and the number of segments belonging to class i, respectively. N is the total number of segments from all classes and n_c is the number of
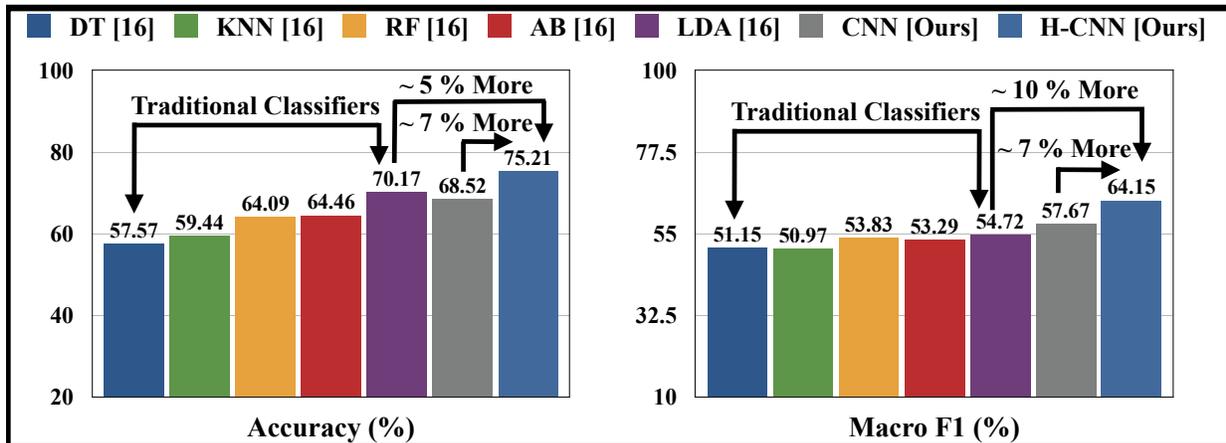
Figure 2.2: Performance Comparison on 3-Class (Baseline vs Stress vs Amusement) Classification

output classes. The *CategoricalCrossentropy* is used as the loss function. We use the *Adam* optimizer with a learning rate of .001. To demonstrate the generalization property of our trained model and to ensure a fair comparison with the traditional classifiers in [40], we also perform Leave One Subject Out (LOSO) validation. As shown in Figure 2.2, the Linear Discriminant Analysis (LDA) classifier in [40] outperforms other classical algorithms for 3-class classification with an accuracy of 70.17% and macro F1 score of 54.72%. Our normal CNN achieves slightly less accuracy of 68.52% compared to LDA but outperforms in macro F1 score with 57.67%. Our H-CNN classifier outperforms both LDA and our normal CNN with an accuracy of 75.21% and macro F1 score of 64.15%. Thus, our H-CNN improves the accuracy by ≈5% and ≈7% compared to LDA and normal CNN, respectively. For macro F1 score, our H-CNN shows higher improvement of ≈10% and ≈7% compared to LDA and normal CNN, respectively. For 2-class (Stress vs. Non-stress) classification, baseline and amusement are considered as the non-stress class. As shown in Figure 2.3, for 2-class classification also, our H-CNN improves the accuracy by ≈3% and ≈5% compared to LDA classifier and normal CNN, respectively. Similarly, for macro F1 score, our H-CNN improves the performance by ≈3% and ≈7% compared to LDA and normal CNN, respectively.
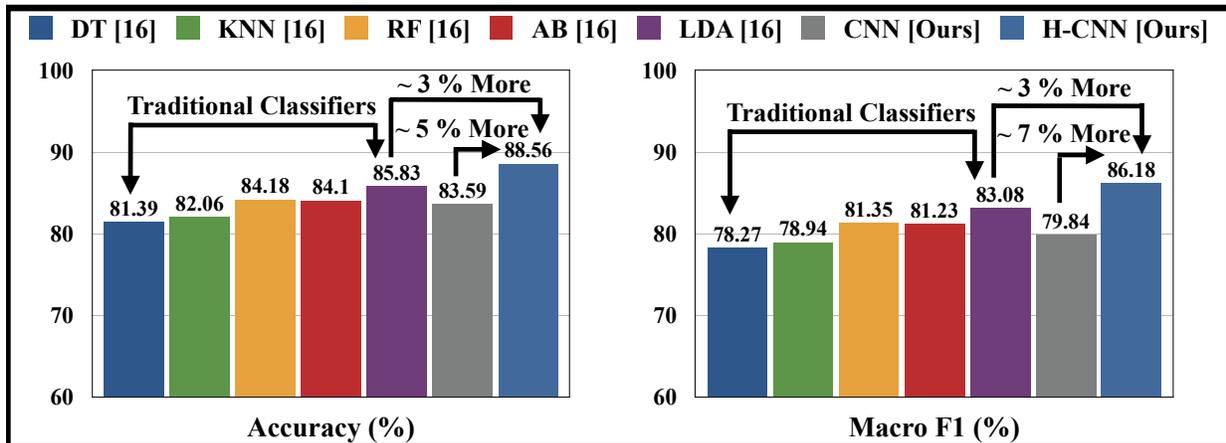
Figure 2.3: Performance Comparison on 2-Class (Stress vs Non-stress) Classification

2.4   Chapter Conclusion

This chapter proposed a novel hybrid CNN (H-CNN) classifier to detect stress using a wrist-based PPG sensor focusing on consumer-grade wristwatches. Our H-CNN uses both the hand-crafted features and the automatically extracted features by CNN to detect stress using the BVP signal. Evaluation on the benchmark WESAD dataset shows that, for 3-class classification (Baseline vs. Stress vs. Amusement), our proposed H-CNN outperforms traditional classifiers and normal CNN by $\approx 5\%$ and $\approx 7\%$ accuracy, and $\approx 10\%$ and $\approx 7\%$ macro F1 score, respectively. Also for 2-class classification (Stress vs. Non-stress), our proposed H-CNN outperforms traditional classifiers and normal CNN by $\approx 3\%$ and $\approx 5\%$ accuracy, and $\approx 3\%$ and $\approx 7\%$ macro F1 score, respectively. To the best of our knowledge, our H-CNN shows the highest performance for both 3-class and 2 -class classification using the BVP signal from the WESAD dataset while performing LOSO validation.

Chapter 3

Neural Contextual Bandits Based Dynamic Sensor Selection for Low-Power Body-Area Networks

3.1 Body-Area-Networks

Machine Learning (ML) based solutions for mobile health applications have shown promising results in providing efficient monitoring of numerous physiological signals such as ECG [28], EEG, or PPG. The increased use of multi-modal devices has led to the introduction of body area networks in wearable devices for continuous real-time monitoring [6]. Traditionally, the main operating principle of these mobile health monitoring devices starts by sending the collected raw data from multiple sensors to an intermediate instrument (e.g., smartphone) using a low-power wireless technology (e.g., Bluetooth). Then, the data is forwarded further to a cloud server for storing or processing heavy algorithms such as Deep Neural Networks (DNN) using long-range energy-hungry wireless methods [44] such as Wi-Fi or LTE. However,

the additional energy associated with raw data transmission over the wireless medium brings concerns about the extra energy consumption of sensor devices for this approach. Moreover, at the end of this transmission, since the redundant data is also sent to the mobile devices, the amount of data that the device is required to process increases, which decreases the device's battery life and makes it challenging to sustain continuous monitoring over long periods. As a result, the current growing trend encourages the adoption of edge computing, where most of the processing is leveraged to the closest devices to data generation. In this setting, the recording device (e.g., smart sensors) can contain a lightweight model to perform the needed processing to decide whether to transmit the useful or most informative part of the data. These lightweight models can be obtained through careful expert design, or through a hardware-software ML design tool like Neural Architecture Search [29, 28] for automated exploration of energy-efficient models. Although additional communication overheads can be discouraging for edge computing, many researchers have shown that by considering the network properties as part of the model design it is possible to minimize the negative effects of communication while still enjoying some of the benefits of edge computing [30, 7].

Different methods exist in the literature regarding filtering out the unnecessary transmission and processing of raw Electrocardiogram (ECG) data to wearable devices. For example, a recent study from Demirel et al. [8] has proposed to use early layers of the neural network to decide whether the current heart activity is normal or abnormal in terms of the beat waveform morphology and heart rate variability; if it is normal, the data is not transmitted to a cloud server for further investigation. However, this approach has not considered the multi-channel element of the ECG signals and focused on identifying small numbers of cardiac abnormalities that do not represent the complexity and difficulty of heart monitoring applications. More recent works [20, 38] have focused on finding the optimal lead subset of the 12-lead ECG at design time to eliminate the redundancy, which can help improve the generalizability of DL-based models and decrease the energy consumption due to communication and classification. However, these proposed methods do not consider dynamic changes in the human heart

while designing their method, which results in a static system. For example, the study [20] investigates what channels are necessary to keep and which ones may be ignored when considering an automated detection system for cardiac ECG abnormalities, and found a 4-lead ECG subset for classification of ECG abnormalities using the DL model. However, in this paper, we showed that 2-lead ECG is also enough for some signal segments, and using a 4-lead ECG would increase the power consumption without increasing the classification performance. Although, the existing literature on ECG abnormality detection is extensive and focuses particularly on multi-lead systems.

Given the aforementioned problems, our proposed methodology contributes in the following way:

- We proposed a dynamic lead selection methodology using neural contextual bandit for high-performance and resource-efficient mobile health applications.

- We proposed a data driven novel contextual bandits environment that uses a simple feature extracted from different sensors in ECG signals to determine which lead combinations can be chosen to decrease communication and computational overhead without sacrificing the classification performance.

- We demonstrate the effectiveness of our proposed methodology in a body area network to monitor cardiac activity.

- On the PTB-XL ECG dataset [48], our proposed methodology achieves state-of-the-art performance metrics for classifying cardiac activity into five diagnoses, reaching 78% AU-PRC while decreasing the overall energy comsumption and computational energy by $3.7\times$ and $4.3\times$ respectively .

15

3.2 Motivation

We have conducted several experiments to show why dynamic lead selection is needed and can help decrease energy consumption significantly while classifying ECG signals with high performance in resource-constrained devices. Firstly, we examine five different classifiers, one for each lead subset (from twelve-lead, six-lead, four-lead, three-lead, and two-lead ECG recordings), to detect normal ECG (NORM), myocardial infarction (MI), conduction disturbance (CD), ST/T-changes (STTC), hypertrophy (HYP) in PTB-XL dataset [48]. The lead subsets are chosen as twelve-lead, six-lead (I, II, III, aVR, aVL, aVF), four-lead (I, II, III, V2), three-lead (I, II, V2), and two-lead (I and II) ECG recordings. While doing the experiments, we followed the suggested evaluation method for the PTB-XL dataset; training on the first eight folds and the ninth and tenth folds are used as validation and test sets, respectively. For each subset and class, the F1 score ($2 \times \frac{Precision \times Recall}{Precision+Recall}$) in percentage is obtained and reported in Figure 3.1. The most interesting aspect of this Figure is that the
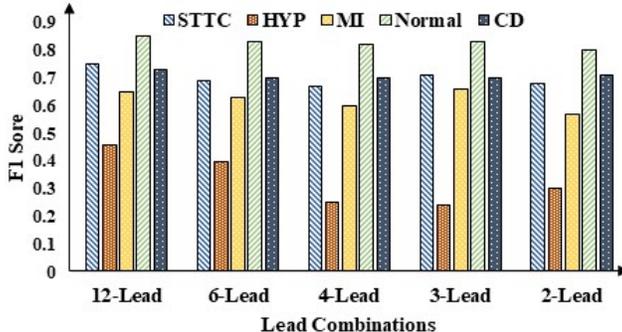
Figure 3.1: The classification performance of each class with different sensor combinations

classification performance of the neural network stays relatively constant, around $80 - 81\%$, for normal ECG (NORM) in each lead combination. However, HYP type segments classification rate decreases heavily in 6 and 3 lead combinations, especially 3-lead is the lowest classification rate with 20% drop compared to the single lead. Although it looks like STTC and CD detection performance stays close to constant in different lead combinations, its percentage is decreased by 7% from the best one in a single lead. Surprisingly, these results show no direct relation with increasing the number of sensors for the classification of different

16

diseases. Also, it should be noted that the classifier designed with less number of sensors, such as 2-lead, has 6× fewer parameters and Floating point operation (FLOPs) compared to the baseline model that uses all available sensors. Moreover, using all available sensors brings additional energy consumption due to transferring raw data to the device for processing. As communication accounts for a significant portion of the total power consumption of a connected device [25], a method to decrease the communication data will enhance battery life significantly and realize the ultra-low power Body-Area Networks. Overall, these results suggest that if the sensor combinations are selected precisely during runtime, significant energy consumption can be prevented without sacrificing the classification performance.

3.3   Methodology

3.3.1   Adaptive CNN Classifier

The designed classifier is a convolutional neural network, which take as input only the raw 10-second ECG segments and no other patient- or ECG-related features, and classifies a single segment into five classes (NORM, MI, CD, STTC, HYP). The architecture is designed to extract various morphological features from the complete segment by employing different lengths of convolutional filters at different layers of the architecture. While designing the architecture, we utilized both the Residual connections introduced by He et al. in [16] and the Inception architecture [45] that has been shown to achieve good performance while maintaining computational and memory costs at low levels. The Conv blocks in Figure 3.2 show the implemented original residual connections where the activation is applied after addition. The model includes three residual blocks with different kernel sizes and filter numbers. Every residual block subsamples its inputs by a factor of 2 by taking the maximum sample (i.e., max pooling with stride 2). The Rectified Linear Unit (ReLu) is utilized as

the activation function in the classifier. All convolutional layers are implemented using a stride of 1 except the first filter, which moves two samples, resulting in half of the samples $K/2$. Unlike the Inception architecture, the wider layer (one after the first residual block) is not stacked up together; instead, we have used Residual blocks, which helps to reduce the dimension of the network while combining the various features of a heartbeat.
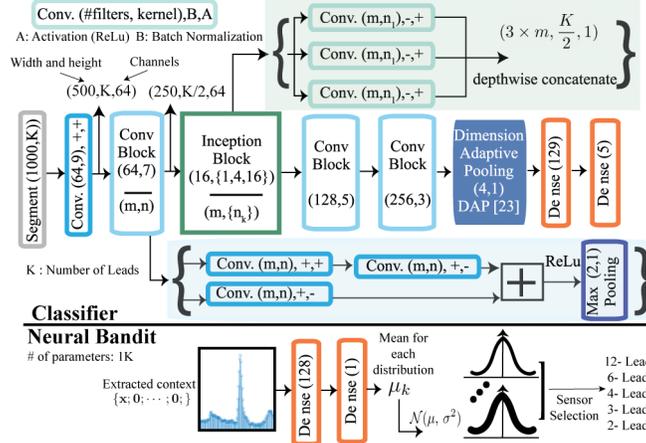
Figure 3.2: Proposed Classifier and Neural Bandit Architecture

Otherwise, the sequential connections of these wider layers result in a quadratic increase of computation and parameters, making the network inefficient and prone to overfitting.

As the number of channels is controlled during runtime, the input to classifier which is represented as $(1000, K, 1)$, can be different. The K value in this representation is the number of sensors which the neural bandits choose dynamically during runtime. However, CNN architectures are generally designed to work with incoming data of a fixed channel size, and changes in the sizes of their inputs cause substantial performance loss, unnecessary computations, or failure in operation. To handle these limitations, we have utilized the dimension-adaptive pooling (DAP) layer [26] that makes DNNs flexible and more robust to changes in sampling rate. After each convolutional layer, we applied batch normalization [17] and/or a rectified linear activation. The '+' and '−' signs near the convolutional layer shows whether these operations are applied for that convolutional operation. We also used Dropout [41] before the last dense layer with a probability of 0.5 to prevent overfitting. The final fully connected sigmoid layer produces a distribution over the five output classes.

18

3.3.2 Neural Thompson Sampling

To apply NTS, we consider the problem of dynamic lead selection as a contextual K-armed bandit problem, where each arm represents a combination of leads and we have a finite number of rounds T. At every round $t \in [T]$, the agent observes K contextual vectors of size d $\{\mathbf{x}_{t,k} \in \mathbb{R}^d | k \in K\}$. When the agent selects an arm a_t it receives a corresponding reward r_{t,a_t}. The goal is to maximize the total expected reward or in other words to minimize the sum of regrets:

$$R_T = \mathbb{E}[\sum_{t=1}^{T} (r_{t,a_t^*} - r_{t,a_t})] \tag{3.1}$$

where a_t^* is the optimal arm at round t which gives the maximum expected reward.

Thompson sampling works by associating each arm with a reward distribution, for simplicity, we use the Gaussian distribution in accordance with [50]. The parameters of distributions are defined through its mean and variance where the trained neural network approximates the mean for each arm. And the variance is approximated using the approximated mean with Equation 3.2. The network is updated every time a reward is observed through gradient descent on the squared loss between the networks predicted reward and the real reward.

$$\sigma_{t,k}^2 = \lambda \mathbf{g}^T(\mathbf{x}_{t,k}; \theta_{t-1}) \mathbf{U}_{t-1}^{-1} \mathbf{g}(\mathbf{x}_{t,k}; \theta_{t-1})/m \tag{3.2}$$

$$\mathbf{U}_t = \mathbf{U}_{t-1} + \mathbf{g}(\mathbf{x}_{t,a_t}; \theta_t) \mathbf{g}(\mathbf{x}_{t,a_t}; \theta_t)^T/m \tag{3.3}$$

Where λ is the regularization parameter, $\mathbf{g}$ is the gradient list, $\mathbf{U}$ is the neural tangent kernel

and m is the width of the neural network.

3.3.3 Context Extraction

We investigated various features of the ECG signals as the context for the neural bandits. In the end, we found that the neural bandits performed the best when using the beat segment as the context. For extracting the beat from 10-second 12-lead signals, we have only used lead-II. Although numerous robust methods have already been available for R peak detection, we have used the Pan-Tompkins algorithm [31], a real-time QRS complex-based heartbeat detection approach that has an accuracy of up to 99.5%, and the algorithm for improving R-peak detection is beyond the scope of this manuscript. Similar to other work [8], we take 0.25 second before the peak and 0.3 second after the peak (totally 0.55 second containing the R peak) to represent the corresponding heartbeat. As the last preprocessing step, the segmented beats are normalized to have a maximum value of 1 before feeding them to the neural bandit.

3.4 Experimental Setup

3.4.1 Training Classifier

We use data from the PTB-XL dataset [48] which comprises 21837 clinical 12-lead ECG records of 10 seconds from 18885 patients, where 52% were male and 48% were female. The ECG statements were assigned to three non-mutually exclusive categories diag (short for diagnostic statements such as "anterior myocardial infarction"), form (related to notable changes of particular segments within the ECG such as "abnormal QRS complex") and rhythm (related to certain changes of the rhythm such as "atrial fibrillation"). In total,

there are 71 different statements, which decompose into 44 diagnostic, 12 rhythm and 19 form statements, 4 of which are also used as diagnostic ECG statements. A hierarchical organization into five coarse superclasses and 24 sub-classes is also provided for diagnostic statements. The interested readers can be referred to the original publication for further details on the dataset, the annotation scheme, and other ECG datasets. In summary, PTB-XL stands out by its size as the to-date largest publicly accessible clinical ECG dataset and through its rich set of ECG annotations and other metadata, which turns the dataset into an ideal resource for the training and evaluation of machine learning algorithms. Throughout this paper, we use the recommended train-test splits provided by PTB-XL [48], which is training on the first eight folds and the ninth, tenth fold is used as validation and test sets, respectively. Also, the input data is used at a sampling frequency of 100 Hz.

Moreover, since the size of each input is controlled during runtime, the CNN classifier should adapt to the changes in the sampling rate. Therefore, we have used adaptive dimension training, which comprises dimension randomization and optimization with accumulated gradients as in [26]. This process works by training the CNN on input data of several randomly selected dimensions (different lead combinations). In this way, the model can learn the waveform morphological features of different sizes of heartbeats.

The classifier network is trained with Glorot initialization of the weights [14]. L2-regularization with 0.0002 is applied for each convolution operation for Inception while the last linear layers are trained with a 0.00005 L2 value. We used the Adam optimizer [19] with the default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a mini-batch size of 80. The learning rate is initialized to 0.001.

3.4.2 Training Bandits

We define the optimal action a_t^*, given a context x_t, as the one which produces the correct classification result and uses the least amount of leads. In the case that none of the lead

combinations can classify correctly, we choose the option with the least amount of leads, in other words the lowest energy consumption. However, it is observed that if the same training data which is employed before for training the ML model is also used for the contextual bandits, the agent becomes highly biased since the optimal arm for each round depends on the correct classification of the current ECG segment and the trained model has already seen those data before in the training phase. This biasing leads to an action space that is highly concentrated in a single action. To prevent this biasing, we construct a mixed training dataset by combining the training and validation where we ignore the majority action samples in the training folds and keep the validation dataset as it is. At the end, it is observed that the combined dataset decreases the biasing in the neural bandits. The training, validation, and test data for the bandit agent follows the same scheme from [48] as mentioned previously in Section 3.4.1.

The contextual bandit neural network is a single hidden layer network with input size equal to K arms times 1000 features, 128 hidden units, and 1 output node to approximate the mean reward. The input contextual vector is constructed following [50] such that each arm sees a corresponding vector in the form $\{\mathbf{x}; \mathbf{0}; \cdots ; \mathbf{0}; \}$ for arm 0, $\{\mathbf{0}; \mathbf{x}; \cdots ; \mathbf{0}; \}$ for arm 1, up to K where x is the feature vector. Two parameters used in NTS are ν and λ for which we set to 1e^{-6} and 1e^{-1} based on a grid search over $\{1, 1e^{-1}, 1e^{-2}, 1e^{-3}\}$ for λ and $\{1e^{-1}, 1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}, 1e^{-6}\}$ for ν. Each context is trained for 100 iterations of stochastic gradient descent with learning rate 1e^{-2} and weight decay $\lambda/counter$ where the counter increases by 1 up until the total training samples.

3.5 Experiments and Results

To evaluate the classification performance, we have used the macroaverage precision, macro-average recall, accuracy, and Area under the Precision-Recall operating characteristic curve

22

(AU-PRC) metrics. The evaluation results of our adaptive performance on PTB-XL dataset is given in Table 2. As shown in the Table, we may conclude that the overall performance of our proposed method overperforms or reaches the baseline model in all leads for ECG classification tasks while being energy and memory efficient.

Table 3.1: Performance Comparison of Proposed method with Related Works

Work	Lead	Precision	Recall	AU-PRC
ML-ResNet [15]	12-Lead	66	**75**	77.9
MFB-CBRNN [23]	12-Lead	65	68	78.6
FCNN [43]	12-Lead	68.1	70.5	77.4
A-CNN [1]	12-Lead	65.8	68.9	77.6
Ours	2-Lead	68	61	71.1
Ours	12-Lead	71	66	78.1
Ours	Adaptive	**72**	63	**78.8**

Especially for the abnormality detection task, the baseline classifier which uses all available sensors (12-leads) performs worse than our proposed adaptive method while increasing the computational complexity of the model more than 6.4×. This shows that using all sensors brings additional energy consumption due to classification and communication while decreasing performance. Moreover, when the proposed method is compared with the related works, it is noticeable that it outperforms or reaches the performance while decreasing the overall calculation for classifiers.

Although the adaptive sensor selection decreases the overall computation and communication while achieving comparable classification performance, it should be emphasized that one of the more significant contributions to emerge from this study is that our proposed method is not a substitute for other methods concerning resource-constrained low-power devices;

instead, it is a complementary method that can be used along with them. For example, hardware-aware hyper-parameter optimization [42] or pruning and quantization [21] of deep learning models are widely used in literature, once those dynamic compression techniques have compressed a network, our adaptive sampling can still be applied to the multi-modal systems and fed to the compressed network. Or, different deep learning architectures which are concerned about energy and memory are recently proposed for ECG classification, our proposed method can be utilized with these architectures as an additional preprocessing step.

3.6 Memory and Energy Consumption Evaluation

We evaluate the proposed method's memory footprint and energy consumption on the EFM32 Giant Gecko ARM Cortex-M3-based 32-bit microcontrollers (MCUs), which has a 1024 kB flash and 128 kB of RAM with CPU speeds of up to 48 MHz. Table 3.2 shows the execution time, energy consumption, and required memory for each operation that runs on the edge device which is deployed to the body area network for selecting sensors during runtime. The operations are implemented and deployed to the target device using MATLAB (MATLAB and Coder Toolbox Release R2021b, The MathWorks, Inc, USA).

Table 3.2: Memory Footprint, Execution Time and Energy Consumption Evaluation on EFM32 Giant Gecko Development Board.

Operations	Exe. Time (ms)	Avg. Energy (μJ)	Flash Memory Footprint (KB)	RAM Memory Footprint (KB)
Context extraction	1200	580.2	9.4	20.7
Post-processing	2.7	1.17	8	43
Neural Bandit	175	8	32.5	35.7
Overall	1377.7	589	$\leq$ 64 KB	$\leq$ 64 KB

The overall execution time for a 10-second ECG segment, which is sampled from the PTB-XL dataset takes 1377.7 ms in the edge device with 589 μJ energy consumption. As the context

extraction operation includes filtering, peak-detection, segmentation and normalization, its computational overhead dominates the overall operations. Also, our proposed method is compatible with any edge device with a minimum RAM of 64 KB. As a result, our method guarantees high performance while maintaining the low-power edge devices' requirements of being resource-efficient in terms of energy and memory. Moreover, We have done several experiments to show the advantages of the proposed method for decreasing the communication energy of the body area networks. The 1-hour raw 12-lead, 2-lead and dynamically selected lead combinations are transferred to the mobile devices without changing the sampling rate of the signal to mimic the working principle of the body area networks. Then, the energy consumption of these transfer operations is calculated for four different communication technologies, *Wi-Fi*, *LTE*, *3G* and *BLE*. The energy consumption for all cases' transfer operation is shown in Figure 3.3.
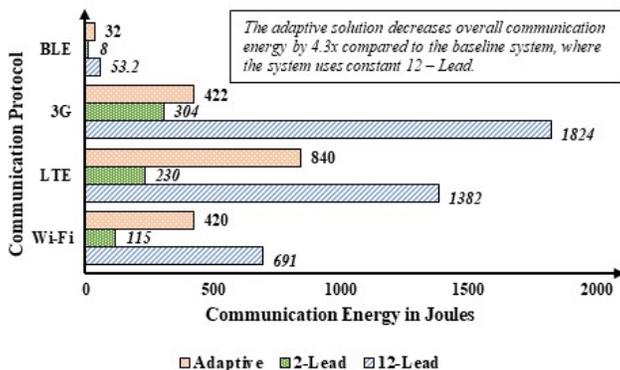
Figure 3.3: The energy consumption of transfer operation for different lead combinations and proposed adaptive solution

While calculating the energy consumption of these cases, we have followed the $\mu J/bit$ values given in [8] for the *Wi-Fi*, *LTE*, and *3G*. For *BLE* protocol energy consumption, we performed the profiling for data exchange on an EFR32BG13 Blue Gecko Bluetooth® Low Energy SoC which has 32-bit ARM Cortex-M4 core with 40 MHz maximum operating frequency. What stands out in Figure 3.3 is that when the proposed adaptive method is utilized instead of using common 12-lead sensor combination in the body area network, the communication energy savings can reach 4.3× in all protocols. Although, the energy consumption of the

2-lead raw data is the lowest, its classification cannot achieve a comparable performance with state-of-the-art works. While comparing the FLOPs, we have only considered the architectures which are given explicitly. Furthermore, any preprocessing such as Fourier or wavelet transforms are ignored, and only the CNN architecture for classification FLOPs is calculated.
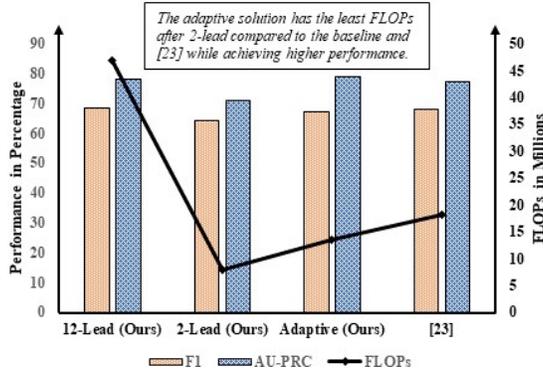
Figure 3.4: Comparison of computational complexity and performance

While calculating the total number of floating-point operations (FLOPs), we have followed [27] where the convolution is assumed to be implemented as a sliding window and that the nonlinearity function is computed for free. For convolutional kernels we have;

$$\text{FLOPs} = 2HW(C_{in}K + 1)C_{out} \qquad (3.4)$$

Where H, W and C_{in} are the height, width and number of channels of the input feature map, K is the kernel width, and C_{out} is the number of output channels. As shown in Figure 3.4, the adaptive solution reduces the overall FLOPs value by 1.35×, and 3.5× compared to [43] and baseline where all available sensors are used. Moreover, while decreasing overall computation, the adaptive solution does not sacrifice any performance even the overall F1 score value increases 10%.

26

3.7   Chapter Conclusion

In this chapter, we proposed a methodology for dynamic sensor selection of body area networks in low-power resource-constrained devices in terms of memory and battery (e.g. wearable devices) using neural contextual bandits. Moreover, we also presented a data-driven novel multi-armed bandits that uses a simple feature extracted from different sensors in ECG signals to determine which lead combinations can be chosen to decrease communication and computational overhead without sacrificing the classification performance. Evaluation of the PTB-XL dataset shows that our proposed dynamic sensor selection solution requires 64 KB of RAM and achieves up to $3.7\times$ and $4.3\times$ overall energy efficiency in computational and communication energy, respectively without sacrificing any classification performance.

Chapter 4

Conclusion

In this thesis, I presented two machine-learning techniques for mobile health applications in low-power embedded systems. The first was a feature-augmented hybrid convolutional neural network architecture for stress monitoring using a wrist-based photoplethysmography sensor. The hybrid-CNN technique uses both the hand-crafted features and the automatically extracted features by CNN to detect stress using the BVP signal. The proposed technique outperformed traditional classifiers and normal CNN by $\approx$5% and $\approx$7% accuracy, and $\approx$10% and $\approx$7% macro F1 score, respectively. Also for 2-class classification (Stress vs. Nonstress), our proposed H-CNN outperforms traditional classifiers and normal CNN by $\approx$3% and $\approx$5% accuracy, and $\approx$3% and $\approx$7% macro F1 score, respectively. The second technique is a neural contextual-bandits-based dynamic sensor selection framework for cardiovascular disease detection with a body-area network of electrocardiogram sensors. Evaluation of the PTB-XL dataset shows that our proposed dynamic sensor selection solution requires 64 KB of RAM and achieves up to 3.7$\times$ and 4.3$\times$ overall energy efficiency in computational and communication energy, respectively without sacrificing any classification performance.

Bibliography

[1] U. R. Acharya et al. Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network. *Information Sciences*, 2017.

[2] A. P. Association et al. Stress in america 2020: A national mental health crisis. *Retrieved December*, 1:2020, 2020.

[3] A. P. Association et al. Stress in america 2021: One year later, a new wave of pandemic health concerns, 2021.

[4] CDC. Nvss public use data file documentation, Dec. 2021.

[5] CDC. Underlying Cause of Death, 1999-2020 Request, 2022.

[6] M. Dautta, A. Jimenez, K. K. H. Dia, N. Rashid, M. A. Al Faruque, and P. Tseng. Wireless qi-powered, multinodal and multisensory body area network for mobile health. *IEEE Internet of Things Journal*, 8(9):7600–7609, 2021.

[7] B. U. Demirel, I. A. Bayoumy, and M. A. A. Faruque. Energy-efficient real-time heart monitoring on edge–fog–cloud internet of medical things. *IEEE Internet of Things Journal*, 9(14):12472–12481, 2022.

[8] B. U. Demirel et al. Energy-efficient real-time heart monitoring on edge-fog-cloud internet-of-medical-things. *IEEE IoT Journal*, 2021.

[9] B. U. Demirel, I. Skelin, H. Zhang, J. J. Lin, and M. Abdullah Al Faruque. Single-channel eeg based arousal level estimation using multitaper spectrum estimation at low-power wearable devices. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 542–545, 2021.

[10] N. Dua, S. N. Singh, and V. B. Semwal. Multi-input cnn-gru based human activity recognition using wearable sensors. *Computing*, 103:1461–1478, 2021.

[11] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[12] S. Elmalaki, B. U. Demirel, M. Taherisadr, S. Stern-Nezer, J. J. Lin, and M. A. A. Faruque. Towards internet-of-things for wearable neurotechnology. In *2021 22nd International Symposium on Quality Electronic Design (ISQED)*, pages 559–565, 2021.

[13] K. Feng, X. Pi, H. Liu, and K. Sun. Myocardial infarction classification based on convolutional neural network and recurrent neural network. *Applied Sciences*, 9(9):1879, 2019.

[14] X. Glorot et al. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. JMLR, 2010.

[15] C. Han et al. Ml-resnet: A novel network to detect and locate myocardial infarction using 12 leads ecg. *Computer methods and programs in biomedicine*, 185, March 2020.

[16] K. He et al. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] S. Ioffe et al. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[18] E. Khalili and B. M. Asl. Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel eeg. *Computer Methods and Programs in Biomedicine*, 204:106063, 2021.

[19] D. P. Kingma et al. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

[20] C. Lai et al. Optimal ecg-lead selection increases generalizability of deep learning on ecg abnormality classification. *Phil. Trans. of the Royal Soc. A: Mathematical, Physical and Engineering Sciences*, 2021.

[21] Y. Li et al. EDD: Efficient Differentiable DNN Architecture and Implementation Co-search for Embedded AI Solutions. In *DAC*, 2020.

[22] W. Lin, B. U. Demirel, M. A. Al Faruque, and G. Li. Energy-efficient blood pressure monitoring based on single-site photoplethysmogram on wearable devices. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 504–507, 2021.

[23] W. Liu et al. Mfb-cbrnn: A hybrid network for mi detection using 12-lead ecgs. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[24] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu. Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In *IJCAI*, pages 3109–3115, 2019.

[25] S. Maity et al. Secure human-internet using dynamic human body communication. In *ISLPED*, 2017.

[26] M. Malekzadeh et al. Dana: Dimension-adaptive neural architecture for multivariate sensor data. *IMWUT*, Sep 2021.

[27] P. Molchanov et al. Pruning convolutional neural networks for resource efficient inference, ICLR 2017.

[28] M. Odema et al. Eexnas: Early-exit neural architecture search solutions for low-power wearable devices. In *ISLPED*, 2021.

[29] M. Odema, N. Rashid, and M. A. Al Faruque. Energy-aware design methodology for myocardial infarction detection on low-power wearable devices. In *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, pages 621–626, 2021.

[30] M. Odema, N. Rashid, B. U. Demirel, and M. A. A. Faruque. Lens: Layer distribution enabled neural architecture search in edge-cloud hierarchies. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 403–408, 2021.

[31] J. Pan and W. J. Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, pages 230–236, 1985.

[32] P. Pimple, B. B. Lima, M. Hammadah, K. Wilmot, R. Ramadan, O. Levantsevych, S. Sullivan, J. H. Kim, B. Kaseer, A. J. Shah, et al. Psychological distress and subsequent cardiovascular events in individuals with coronary artery disease. *Journal of the American Heart Association*, 8(9):e011866, 2019.

[33] E. Prabhakararao and S. Dandapat. Myocardial infarction severity stages classification from ecg signals using attentional recurrent neural network. *IEEE Sensors Journal*, 20(15):8711–8720, 2020.

[34] N. Rashid and M. A. Al Faruque. Energy-efficient real-time myocardial infarction detection on wearable devices. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4648–4651. IEEE, 2020.

[35] N. Rashid, B. U. Demirel, and M. A. Al Faruque. Ahar: Adaptive cnn for energy-efficient human activity recognition in low-power edge devices. *IEEE Internet of Things Journal*, 2022.

[36] N. Rashid, B. U. Demirel, M. Odema, and M. A. Al Faruque. Template matching based early exit cnn for energy-efficient myocardial infarction detection on low-power wearable devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(2), jul 2022.

[37] N. Rashid, T. Mortlock, and M. A. Al Faruque. Self-care: Selective fusion with context-aware low-power edge computing for stress detection. In *2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 49–52. IEEE, 2022.

[38] M. A. Reyna et al. Will Two Do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. *Computing in Cardiology*, 2021.

[39] S. M. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. H. Chon. A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photo-plethysmogram sensor. *Sensors*, 16(1):10, 2015.

[40] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[41] N. Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 2014.

[42] D. Stamoulis et al. Hyperpower: Power- and memory-constrained hyper-parameter optimization for neural networks. In *DATE*, 2018.

[43] N. Strodthoff et al. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological Measurement*, Jan 2019.

[44] M. Sun et al. Wireless power transfer and data communication for low-power micro electronic devices deeply implanted within the human body. In *ISLPED*, 2021.

[45] C. Szegedy et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[46] V. Vaccarino, Z. Almuwaqqat, J. H. Kim, M. Hammadah, A. J. Shah, Y.-A. Ko, L. Elon, S. Sullivan, A. Shah, A. Alkhoder, et al. Association of mental stress–induced myocardial ischemia with cardiovascular events in patients with coronary heart disease. *JAMA*, 326(18):1818–1828, 2021.

[47] V. Vaccarino, S. Sullivan, M. Hammadah, K. Wilmot, I. Al Mheid, R. Ramadan, L. Elon, P. M. Pimple, E. V. Garcia, J. Nye, et al. Mental stress–induced-myocardial ischemia in young patients with recent myocardial infarction: sex differences and mechanisms. *Circulation*, 137(8):794–805, 2018.

[48] P. Wagner et al. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 2020.

[49] J. Wei, C. Rooks, R. Ramadan, A. J. Shah, J. D. Bremner, A. A. Quyyumi, M. Kutner, and V. Vaccarino. Meta-analysis of mental stress–induced myocardial ischemia and subsequent cardiac events in patients with coronary artery disease. *The American journal of cardiology*, 114(2):187–192, 2014.

[50] W. Zhang et al. Neural thompson sampling. In *The International Conference on Learning Representations, ICLR*, 2021.