Novel Processing and Transmission Techniques Leveraging Edge Computing for Smart Health Systems

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Electrical, Electronics and Communications Engineering
(31^{th}cycle)

Novel Processing and Transmission Techniques

Leveraging Edge Computing for Smart Health Systems

By

Alaa Awad Abdellatif

******

Supervisor(s):
Dr. Amr Mohamed, Supervisor
Dr. Carla-Fabiana Chiasserini, Co-Supervisor

Doctoral Examination Committee:
Prof. Hossam S. Hassanein, Referee, Queen's University
Prof. Ridha Hamila, Referee, Qatar University
Prof. Mohsen Guizani, Qatar University
Prof. Emilio Leonardi, Politecnico di Torino
Dr. Alessandro Nordio, Politecnico di Torino

Politecnico di Torino
2018

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Alaa Awad Abdellatif
2018

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Acknowledgements

This work was made possible by GSRA grant # GSRA2-1-0609-14026 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

Abstract

The growing number of chronic disease patients, emergency and disaster management, that demand constant monitoring and examination of human biosignals and vital signs, have increased the prominence of telemonitoring systems. There is also a top national interest worldwide to reduce costs of healthcare services while maintaining high-quality care to patients. Coinciding with this worldwide interest, the rapid advances in Edge Computing, wireless communication technologies, Internet of Things (IoT), and Big Data have facilitated the development of smart-health (s-health) systems. S-health systems leveraging the wide range of technologies (e.g., smartphones, wearable devices, and portable health devices) enable providing efficient continuous-remote healthcare services. However, the need of delivering decent healthcare services to the patients while reducing healthcare costs is a challenging issue. S-health systems require recording, transmitting and processing large volumes of multimodal medical data generated from different types of sensors and medical devices, which is challenging and may turn some of the remote health monitoring applications impractical. One of the promising approaches for enabling s-health is adopting edge computing capabilities with next generation wireless networking technologies to provide real-time and cost-effective healthcare services.

In this thesis, we present our vision for the benefits of exploiting multi-access edge computing (MEC) within the field of s-health. We envision a MEC-based architecture and discuss the benefits that it can bring to realize context-aware approaches so that the s-health requirements are met. In particular, we propose four main approaches that can be implemented leveraging such an architecture to provide efficient data delivery, namely, adaptive data classification and compression at the edge, data-specific transceiver design for healthcare applications, distributed in-network processing and resource optimization, and dynamic networks association. The first approach allows for efficient and low distortion compression, while ensuring high-reliability and fast response in case of emergency applications lever-

aging fuzzy classification and feature extraction techniques. The second approach proposes an efficient transceiver design that reduces amount of transmitted data, while considering the characteristics of the acquired data as well as maintaining application Quality-of-Service (QoS) requirements. The third approach enables data transfer from mobile edge nodes to the cloud in an energy-efficient and cost-effective manner leveraging available network resources and applications' characteristics. The fourth approach focuses on how to benefit from the integration of multiple Radio Access Technologies (RATs) within the MEC architecture, in order to meet the applications' requirements, and optimize medical data delivery. Finally, we discuss several opportunities that edge computing can facilitate for s-health to inspire more research in this direction.

Contents

List of Figures

List of Tables

Acronyms

AANS	Autonomous Access Network Selection
AC	Active seizure
ANSC	Adaptive Network Selection and Compression
BASN	Body Area Sensor Network
CA	Classification Accuracy
CbC	Class-based Compression
DOA	Distributed Optimization Algorithm
DWT	Discrete Wavelet Transform
ECG	Electrocardiography
EEG	Electroencephalography
FCA	Formal Concept Analysis
FD-FE	Frequency-Domain Feature Extraction
IoMT	Internet of Medical Things
MEC	Multi-access Edge Computing
MHC	Mobile-Health Cloud
NAC	Non Active seizure
PDA	Personal/Patient Data Aggregator
RAN	Radio Access Network
RM	Remote Monitoring
RNS	Ranked Network Selection
SBC	Stream-Based Compression
SF	Seizure Free
s-health	smart health
SIC	Swift In-network Classification
TBC	Threshold-based Compression
TD-FE	Time-Domain Feature Extraction
UBA	Uniform Bandwidth Allocation

Chapter 1

Introduction

The investment in the healthcare field becomes a top national interest worldwide. The governments have a moral obligation to provide decent healthcare services for the chronically ill and elderly people [2]. However, traditional healthcare systems cannot provide the scalability needed to cope with the increasing number of elderly and chronic disease patients, given that they require one-to-one relationship between the caregiver and the patient. Hence, the evolution of computational intelligence systems and Internet of Medical Things (IoMT), along with the advances of next-generation wireless technologies, has boosted the development of traditional healthcare processes into smart health services.

One of the promising approaches for enabling smart health services is to have patients participate in their own treatment by providing them with intuitive, non-intrusive tools that allow them to communicate efficiently with their caregivers. Thus, smart-health (s-health) systems have emerged to provide healthcare stakeholders with innovative techniques and tools that promote novel approaches to acquire, process, transfer, and store the medical data. Part of the s-health concept is remote health monitoring, where patients and caregivers can leverage wireless technologies for transferring healthcare information without physical contact. S-health can be considered as the context-aware evolution of mobile health systems [3]. In this thesis, context aware refers to exploiting medical devices and wireless communication technologies to detect the patient context, i.e., patient conditions, gathered data characteristics, and surrounding networks state. This allows us to design intelligent techniques that significantly improve the scalability of such health systems through

optimizing the delivery of the patient's data from the edge network to the cloud. We argue that moving computational intelligence to the network edge is of paramount importance to provide efficient and convenient ways for continuous-remote monitoring. Realizing such concept can assist in reducing hospitalization and enable timely delivery of healthcare services to distant communities at low costs [4]. However, critical challenges have emerged, in such systems, that need to be addressed. In what follows, we will discuss these challenges along with the requirements of relevant e-health systems; then we will introduce our system architecture and thesis objectives to tackle these challenges and meet s-health requirements.

1.1 Motivation

S-Health applications are expected to inspire substantial evolutions in the healthcare industry toward Healthcare Industry 4.0 (Health 4.0) [5], especially in pre-hospital emergency situations and for chronic disease monitoring. Health 4.0 facilitates the automation and personalization of the whole medical process by leveraging medical cyber-physical systems, IoMT, and Edge/Cloud computing. Furthermore, it allows, on one hand, patients to monitor their health without the frequent visits of the hospital or clinic. On the other hand, the healthcare service providers can provide patients with medical services through computerized medical information systems and efficient bidirectional communication. However, new challenges have emerged with these major trends including:

- The massive real-time data collected by health monitoring systems. In s-health systems, besides IoMT, various wireless sensors, cameras, and controllers play an important role: they allow patients' automatic identification and tracking, correct drug patient associations, and continuous real-time vital signs monitoring for early detection of clinical deterioration (e.g., seizure detection, heart failure, fall detection, etc.). All these sources report an enormous amount of real-time data that needs to be transported, swiftly processed, and stored.[1]

- Limited power sources and small form factor of the sensor nodes. E-health systems typically consist of several battery-operated devices that should run

[1] It is expected that data produced by people, machines, and IoT will reach 500 zettabytes by 2019 [6].

for a long time without replacement. Hence, continuous data transmission is not viable due to the high energy toll it implies.

- The need for fast prediction and detection of emergency situations. The increased response time due to large volume of data to be delivered and processed may threaten people's lives.

For the sake of completeness, we highlight the above challenges through discussing some of the relevant e-health systems in the next section.

1.2 Characteristics and Requirements of the Relevant E-health Systems

In Table 1.1 and Table 1.2, we summarize the characteristics and requirements of some of the e-health systems. It is not the objective of this thesis to provide an in-depth technical comparison on the different proposed e-health systems. However, we investigate the practical challenges of s-health applications in the light of these systems.

1.2.1 Monitoring systems using wearable devices

Heart monitoring applications are the most common type of remote monitoring applications. Monitoring vital signs related to the heart reveals many types of diseases, e.g., Cardiac arrhythmia, chronic heart failure, Ischemia and Myocardial Infarction (MI) [7][8][9]. In [7], authors present a real-time heart monitoring system, where the extracted medical data of the patients are transmitted to an Android based listening port via Bluetooth. Then, this listening port forwards these data to a web server for processing. Also, [8] exploits Android smartphone to gather patient's information from wearable sensors and forward it to a web portal in order to facilitate the remote cardiac monitoring. However, in these systems, the smartphone is used only as a communication hub to forward collected data to the cloud.

There is also a prompt progress in the field of neurologically-oriented monitoring systems. However, the enormous amount of data generated from such systems is also challenging. For instance, in Intensive Care Unit (ICU) of Electroencephalography

Table 1.1 Sample of the e-health monitoring systems.

Application	Description	Limitations
Cardiac disorder detection [7]	Heart monitoring system is proposed for detecting patients' state and sending an alert message in case of abnormalities *Data*: Electrocardiography (ECG) *Requirements*: long lifetime for the battery-operated devices	All data processing tasks are performed at a web server
Remote Cardiac monitoring [8]	A location based real-time Cardiac monitoring system is developed *Data*: Heart rate, blood pressure, and body temperature *Requirements*: long lifetime for the battery-operated devices	Few number of subjects participated in the experiments
Detection of Ischemia and Myocardial Infarction [9]	Different methods to detect Ischemia and MI are presented using ECG signal along with Electronic Health Records (EHR) information *Data*: ECG and EHR *Requirements*: low computational complexity	Majority of the reviewed methods did not exploit contextual information
Parkinson's disease (PD) detection [10]	A PD monitoring system over the cloud is proposed using feature selection and classification *Data*: Voice signal *Requirements*: Reliability and high classification accuracy	All data processing tasks are performed at the cloud
Contactless heart rate measurement [11]	Heart rate measurements from facial videos are performed using digital camera *Data*: Heart rate *Requirements*: Reliability and high measurement accuracy	Illumination variance, motion variance, and motion artifacts

(EEG) monitoring system, samples of EEG along with video recording should be stored and accessed remotely for correlating clinical activity with EEG pattern. This can result in generating 8-10 GB per patient every day [12], which obviously sets a significant load on the system design and scalability in terms of processing capabilities, storage space, and transmission power.

1.2.2 Contactless monitoring systems

Along with the evaluation of remote sensing, contactless monitoring has attained much focus recently. The main motivation of using contactless sensors is enabling ordinary life as much comfortable as possible to all patients, since the patients are required only to be present within a few meters from the sensors [10]. Heart rate measurement from facial videos using digital camera sensors is one of the rapidly growing directions to extract physiological signals without affecting patient's activities [11]. However, transmitting large volumes of data generated from these camera sensors using conventional cloud-based architecture is not advisable and may deem some of these applications impractical given the limited network resources. For instance, the amount of digital data generated from a single-standard camera can reach to 40 GB per day.

1.2.3 Disorder prediction/detection systems

One of the promising applications of e-health systems, is the predictive monitoring of high-risk patients. The aim of these techniques is improving prediction/detection of the emergency to implement preventative strategies for reducing morbidity and mortality associated with high-risk patients. For instance, [13] presented a simplistic framework for near-term prediction of Bradycardia in preterm infants using statistical features extracted from ECG signal. Also, [14] proposed a quick seizure detection algorithm using fast wavelet decomposition method. In such real-time prediction/detection systems, the swift delivery of data to the server is a necessity. In many cases, this requires that data are analyzed and possibly diagnosis is made as close as possible to the patient in order to reduce the response time. However, detecting the changes of the physiological signals (e.g., abnormality in ECG signals) in continuous health monitoring systems is not an easy task. It can be an indication for an emergency situation (e.g., occurrence of a heart attack) [15][16]. This

abnormality detection task becomes even more challenging during wireless communication transfer of patient's data to the cloud due to the erroneous communication and security attacks that could introduce errors or affect patient data integrity.

Such requirements of e-health systems make the conventional cloud computing paradigm unsuitable for s-health, since the centralized approach cannot provide a sufficiently high level of scalability and responsiveness while causing heavy network load.

Table 1.2 Sample of the e-health prediction/detection systems.

Application	**Description**	**Limitations**
Prediction of Bradycardia in preterm infants [13]	Leveraging point process analysis of the heartbeat time series for near-term prediction of Bradycardia in preterm infants *Data*: ECG *Requirements*: Fast prediction of emergency situations	Using single channel ECG data to predict Bradycardia
Real-time epileptic seizure detection [14]	Automatic epileptic seizure detection system is developed using wavelet decomposition *Data*: EEG *Requirements*: Fast seizure detection	Requiring large amount of data for training to improve specificity of the detector
ECG change detection [15]	A centralized approach for the detection of abnormalities and intrusions in the ECG data is developed *Data*: ECG *Requirements*: Fast detection of abnormalities	Using one type of data for detecting abnormality and emergency situations
Remote monitoring of chronic obstructive pulmonary [16]	Real-time tracking system of chronic pulmonary patients comfortable in their home environment is developed *Data*: Pulmonary Function Test (PFT) *Requirements*: Fast detection of abnormalities	Relying on one type of data

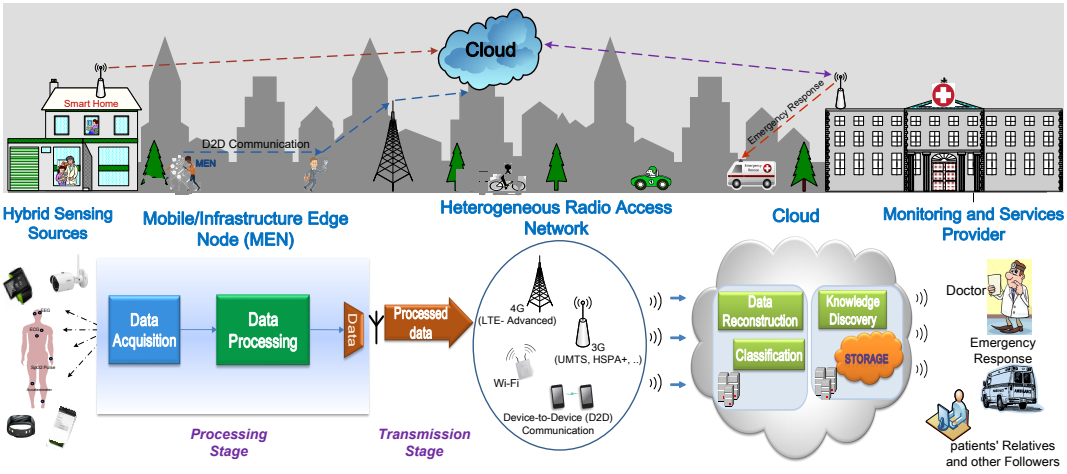


Fig. 1.1 Proposed smart health system architecture.

1.3 Multi-access Edge Computing for Smart Health

Edge computing can significantly help in the healthcare evolution to smart healthcare through enabling better insight of heterogeneous healthcare environment in order to provide affordable and quality patient care [17]. Hence, a new approach has emerged, known as Mobile Edge Computing or Multi-access Edge Computing (MEC), defined as the ability to intelligently process data at the edge of the network, i.e., at the proximity of the data sources [18, 19]. Thus, this section gives a brief description of our MEC-based architecture for s-health systems (see Figure 1.1), while discussing the benefits that it offers to such systems given the above challenges.

1.3.1 MEC-based Architecture

The presented system architecture, shown in Figure 1.1, stretches from the data sources located on or around patients to the healthcare service providers. It contains the following major components:

1) Hybrid sensing sources: A combination of sensing devices attached/near to the patients represents the set of data sources. Examples include: body area sensor networks (including wearable medical and non-medical sensors), IP cameras, smartphones, and external medical devices. All such devices are leveraged for

monitoring patients' state within the smart assisted environment, which facilitates continuous-remote monitoring and automatic detection of emergency conditions.

2) Mobile/infrastructure Edge Node (MEN): Herein, a MEN can be a mobile edge node, which works as a Personal/Patient Data Aggregator (PDA), or an infrastructure edge node that is deployed near to the patient to aggregate the data collected by a Body Area Sensor Network (BASN) and transmit it to the cloud. This MEN implements intermediate processing functions between the data sources and the cloud. In particular, the MEN fuses the medical and non-medical data from different sources, performs in-network processing on the gathered data, classification, and detection of the patient context, extracts information of interest, and forwards the processed data or extracted information to the cloud. In addition to that, the MEN can be a data source itself, which acquires signals related to the health of the patients as well as context data. Importantly, various e-health applications (apps) can also be implemented in the MEN, e.g., for long-term chronic disease management. Such apps can help patients to actively participate in their treatment and to ubiquitously interact with their doctors anytime and anywhere. Furthermore, with a MEN running specialized context-aware processing, various data sources can be connected and managed easily near the patient in order to optimize data delivery based on the patient's context, i.e., data type, supported application, and wireless network conditions.

3) Heterogeneous Radio Access Network: Several s-health services demand for high data rates and quality of service (QoS) level, thus motivating the use of wireless heterogeneous network (HetNet). Thanks to the availability of different technologies such as UMTS (Universal Mobile Telecommunications Service), LTE (Long-Term Evolution), Wi-Fi, and Bluetooth, HetNet can face the increasing traffic demand and successfully meet the QoS requirements. Also, it allows the selection of the most convenient radio interface, in terms of energy as well as monetary cost.

4) Cloud: It is a healthcare cloud where data storage, sophisticated data analysis methods for pattern detection, trend discovery, and population health management can be enabled. An example of the healthcare cloud can be a hospital, which monitors and records patients' state while providing required help if needed.

5) Monitoring and services provider: A health service provider can be a doctor, an intelligent ambulance, or even a patient's relative, who provides preventive, curative, or emergency healthcare services to the patients.

1.3.2 Benefits of MEC-based Architecture for S-health

In the light of the aforementioned challenges described in Section 1.2, the proposed MEC-based system architecture exhibits the following major advantages:

Reduced amount of transferred data: E-health monitoring systems require the collection, hence transmission, of large volumes of data (as shown in Table 1.1), since many sensors and medical devices continuously generate a massive amount of data every few seconds [4]. Thus, considering centralized cloud to support such traffic workload is not advisable and may deem some of these applications impractical given the limited bandwidth availability and battery-operated devices. On the contrary, processing, compressing, and extracting most important information from the gathered data at the MEN greatly reduce the amount of data to be transferred toward the cloud, hence reduce the bandwidth consumption, and even makes it possible to store the data locally.

Extended devices lifetime: Given the requirement of e-health monitoring systems discussed in Section 1.2.1 and 1.2.2, managing the devices operational state and their data transfer at the MEN allows for a better usage of the devices' battery. Moreover, the proximity between devices and MEN further reduces the energy consumption due to data transmission. Regarding the data transmission from the MEN toward the cloud, data compression and selection of the most convenient radio interface to be used, can also significantly decrease the energy consumption at the MEN – a component that amounts to about 70% of the total power consumption of a wireless monitoring system [20].

Low latency: For real-time high-intensive monitoring applications, the swift delivery of data to the cloud is a necessity (as shown in Table 1.2). In many cases, this requires that data is analyzed and possibly diagnosis is made as close as possible to the patient. Hence, quick detection of the changes in the gathered medical data at the MEN can significantly help in real-time abnormal event detection. The implementation of a smart edge node addresses this issue, and the ability of the MEN to perform event detection fulfills this requirement even in the case of emergency applications.

Location Awareness: Given the requirements of prediction/detection systems discussed in Section 1.2.3, the network edge can be fruitfully exploited to extract context information and apply localization techniques. This brings two main ad-

vantages to s-health applications. First, localizing a patient allows matching his/her geographical position with the nearest appropriate caregivers (e.g., hospital or ambulance). Second, data delivery can be optimized accounting for the nearest edge node, or the most suitable device that can further relay data to the network infrastructure, which ultimately lowers the total energy consumption and increases reliability.

1.4 Thesis Objectives and Contributions

This thesis presents a generalized s-health architecture for reliable, scalable, and effective patient monitoring, leveraging sensors and smartphone technologies for connecting patients' networks with medical infrastructure to facilitate remote diagnosis and treatment. In particular, we propose an edge-based s-health framework for improving healthcare services through the development of smart techniques, algorithms, and tools enabling the early diagnosis, remote monitoring, and fast emergency response for the elderly and chronic disease patients. In contrast to the previous work in this domain, the adopted framework considers context-aware approaches by focusing on applications' requirements and patients' data characteristics, leveraging heterogeneous wireless networks for optimized medical data delivery. Accordingly, we focus in this thesis on answering the following major questions:

1. How to reduce the amount of transmitted data, while providing reliable healthcare services?
2. How to integrate wireless network components, and application-layer characteristics to propose energy-efficient s-health system?
3. How to leverage the spectrum across multiple RANs, in order to maintain healthcare applications' QoS?

1.4.1 Objectives

Given the aforementioned challenges and requirements for e-health systems, our thesis targets the following major objectives:

(A) Design a scalable s-health architecture that enables the development of processing and event-detection solutions for collecting, processing, and transferring

patients' data (e.g., vital signs, and health information) to healthcare service providers in an energy-efficient manner.

(B) Develop optimized transmission mechanisms over heterogeneous s-health system leveraging networks characteristics and applications' requirements, while addressing the tradeoff between energy consumption and latency.

1.4.2 Contributions

In order to achieve the above objectives, we propose different context-aware techniques over the proposed MEC architecture (see Figure 1.2). These techniques can: promote effective and convenient ways for continuous-remote monitoring, foster supportive environments for enhancing efficiency of patient treatment, enhance overall system performance, and satisfy the s-health requirements. Specifically, our main contributions can be summarized as follows:

1. Develop automatic, fast-response, and highly-scalable s-health system, which aims to reduce unnecessary doctor's visits through enabling the early diagnosis and detection of the health risks. In our way to maintain objective (A), adaptive edge-based classification and compression techniques are proposed in the *Processing Stage* to be implemented in the devices nearest to the patients and infrastructure components, e.g., gateways, routers and access points (see Figure 1.1). These techniques could significantly improve medical data delivery and QoS for s-health systems through moving intelligence to the network edge while accounting for the importance and characteristics of the gathered data, as well as patient's state. Description of the proposed techniques can be found in Chapter 2.

2. Design a data-specific low-complexity transceiver that maintains application QoS requirements, i.e., signal distortion, taking into consideration the characteristics of the transmitted data, while saving a significant amount of data that needs to be transmitted. In particular, in the *Transmission Stage* , we propose an efficient compression scheme at the physical layer exploiting the existing Orthogonal Frequency Division Multiplexing (OFDM) transceiver's components in order to obtain high compression efficiency without adding

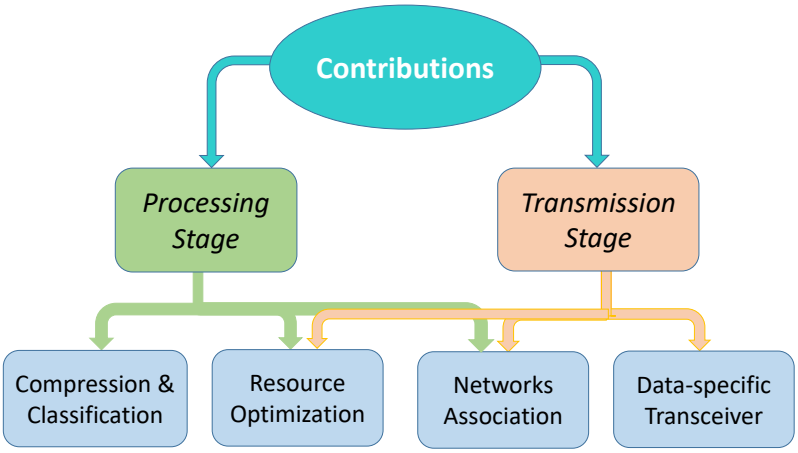


Fig. 1.2 The main contributions of the thesis.

much overheads, as illustrated in Chapter 3. This work helps in achieving objective (B).

3. Formulate and analytically solve a multi-objective optimization problem that targets optimizing different QoS metrics, namely, signal distortion, delay, and Bit Error Rate (BER), as well as monetary cost and transmission energy at the processing and transmission stages. Specifically, we aim to achieve both objectives (A) and (B) through obtaining the optimal tradeoff among the above factors, which exhibit conflicting trends. Thus, in Chapter 4, we propose an optimal, centralized solution leveraging geometric program transformation and Lagrangian duality theory, in addition to a scalable distributed solution. The proposed distributed algorithm converges to the optimal solution and adapts to varying network conditions. Hence, it enhances the overall system performance while providing energy-efficient and reliable connectivity over s-health system.

4. Develop efficient networks association mechanism with adaptive compression for improving medical data delivery over heterogeneous s-health systems. Thanks to the availability of several cellular, WiFi and fixed access technologies, the performance of s-health systems can be significantly enhanced by enabling efficient data transfer from edge nodes to the cloud in an energy-efficient manner, while maintaining the applications' requirements for low latency. This calls for innovative networks association mechanisms at the edge,

which provide optimized data delivery over heterogeneous s-health systems, hence achieving objective (B). Thus, in Chapter 5, we leverage optimization techniques for networks association to allow the edge nodes to minimize their energy consumption and monetary cost, while considering the different QoS constraints, dynamics and characteristics of the available Radio Access Networks (RANs), and applications' requirements.

Chapter 2

Edge-based Compression and Classification for Real-time Healthcare Systems

2.1 Overview

This chapter presents our vision of leveraging edge computing to monitor, process, and make autonomous decisions for smart health applications.[1] It focuses on remote health monitoring system for brain disorders and, in particular, it proposes an energy-efficient remote monitoring system for epileptic seizure detection.[2] EEG signal plays an important role in the diagnosis of epileptic disease, brain death, tumors, stroke, and several brain disorders [24]. Such applications typically require the recording, transmission, and processing of very large volumes of data. Consider, for instance, high-resolution EEG devices consisting of up to 100 electrodes, each working at sampling rate as high as 1000 samples/s. By representing each sample by 2 bytes, it results in a data rate of 1.6 Mbps per single patient. Also, in normal conditions, information about medical patients should be reported to the Mobile-Health Cloud (MHC) every 5 minutes, while, in the case of emergency where high-intensive monitoring is needed, all data collected by the BASN should be reported every 10

[1] This work has been published in [21][22].

[2] Epilepsy is the most common neurological disorder in the world after stroke and Alzheimer's disease. It is estimated to affect more than 65 million people worldwide, with more than 80% of people with epilepsy living in developing countries [23].

seconds [25]. Furthermore, the wireless transmission of such amount of data is highly energy consuming (it amounts to about 70% of the total power consumption of a wireless EEG monitoring system [20]); also, it requires significant processing capabilities, high reliability and, in the case of emergency, very short latency.

Such requirements cannot be supported by resource-constrained PDA, unless we adopt smart solutions. The conventional remote monitoring system using simple sensor-to-cloud architecture [26], where the raw date is collected from different sensor nodes and send to the cloud for processing, becomes unsuitable for s-health. Such centralized approach cannot provide sufficient scalability and responsiveness, while causing heavy network load. On the contrary, by leveraging edge computing capabilities, s-health systems can significantly improve medical data delivery while decreasing the latency and energy consumption.

Given the aforementioned requirements and constraints of remote monitoring systems, our goal is to enable energy-efficient delivery of real-time medical data through implementing: (i) a mechanism for abnormal pattern detection at the network edge that allows us to identify the patient's state, and (ii) a selective transfer scheme that, exploiting the above detection mechanism, transmits toward the MHC only the essential data based on the current situation. Hence, our main contributions can be summarized as follows.

1. Design an energy-efficient remote monitoring system for epileptic seizure detection and notification, which adapts the type of information to be transmitted over the wireless channel based on the patient's state. In the proposed system, local in-network processing at the edge is executed on the raw EEG data before their transmission in order to accurately detect patient's state. Then, depending on the patient's state, our system leverages different data reduction techniques to reduce the amount of transmitted data.

2. Develop a fuzzy classification technique based on feature extraction and Formal Concept Analysis (FCA). Our classifier, named Swift In-network Classification (SIC), enables an accurate detection of the patient's state, while providing a quick notification about the patient's state at the PDA, as well as at the remote monitoring server (doctor's machine). We remark that the proposed mechanism allows for a quick response while keeping the complexity low, thus it is amenable for implementation at the PDA (i.e., mobile edge).

3. Present a comparative study of the applied time-domain and frequency-domain feature extraction techniques, discussing the tradeoff that they exhibit in terms of transmitted data length and classification accuracy.

4. Propose an automated class-based compression scheme that allows for satisfying application QoS requirements (i.e., signal distortion and classification accuracy), while saving a significant amount of energy at the edge considering the class of the data.

5. Evaluating the proposed mechanisms using an implemented real-time EEG monitoring system. In this implemented system, the PDA handles the EEG readings via a specific mobile application that we have developed, and applies proposed SIC technique for detecting the patient's state. Based on the detected state, the PDA transmits the appropriate data type to a remote server, while in case of emergency, an emergency notification is declared at the PDA and forwarded to the server. At the server, the transmitted data from the PDA is received, and a real-time data reconstruction and distortion evaluation for the compressed-received data is applied.

2.2 Related Work

The rapid growth of IoMT has motivated the development of innovative applications for information intensive fields such as healthcare services [27]. The conventional cloud computing architecture facilitates for the smart devices (e.g., sensors, smartphones) to exchange information with the cloud through 3G/4G technologies, or IoT gateway [28]. Thus, on one hand provides uniform, concise, and scalable processing as well as storage services for supporting application requirements. On the other hand, the deployments of remote monitoring applications, and in general delay-sensitive IoT applications on the cloud are facing challenges. For instance, the delay caused by transferring data to and from the cloud is unpredictable, in addition to the economic considerations, technical limitations, and administrative issues [29].

In [30], the authors implement an automatic mobile-based health system exploiting the information contained in EEG signals for seizures detection. This system consists of back-end part (i.e., server part) and front-end part (i.e., mobile part). Server part comprises the pre-processing tasks, which include feature extraction,

normalization, and selection, as well as classification task. While mobile part includes data acquisition, visualization, and transmission. The authors in this study also present different algorithms in the pre-processing and classification stages to implement a reliable system in terms of execution time and classification accuracy. However, they consider the mobile part as a communication hub, while moving the pre-processing and classification tasks to the server. Considering such centralized approach cannot provide a sufficiently high level of scalability and responsiveness given the limited bandwidth availability, energy consumption, and data privacy concerns. On the contrary, processing and compressing the gathered data at the edge greatly reduce the amount of information to be transferred toward the cloud, hence the bandwidth and energy consumption, while ensuring privacy protection.

To address the aforementioned challenges, Edge computing and Fog computing were proposed to use computing resources near IoMT devices for local storage and preliminary data processing [31]. In this context, the authors in [32] proposed a software framework for healthcare applications based on Edge paradigm. This framework is used for acquiring and analyzing Heart Rate Variability (HRV) signals, while presenting the advantages of leveraging Edge paradigm rather than the classical Cloud paradigm. The authors in [33] discuss the benefits of Fog architecture in preventive healthcare applications, and the feasibility of Field-Programmable Gate Array (FPGA) technology in implementing efficient Fog nodes. This aforementioned work motivates that performing efficient in-network processing with feature extraction and adaptive compression at the edge would significantly assist in network congestion, offload core network traffic, accelerating analysis, and meeting application requirements for swift and secure data transfer.

Most of the related work in the context of Wireless Sensor Network (WSN) was motivated by the reduction of latency to perform classification at the sensor network. Different machine learning methods have been investigated to exploit historical data and improve the performance of sensor networks through discovering important correlations in the sensor data and adopting better sensor deployment for maximum data coverage. For instance, the authors in [34] present a classification technique for efficient data collection in WSN. However, it is assumed that the end users are interested only in rounds of measurements characterized by certain patterns. Hence, the WSN exploits the classification with the goal of selecting the most relevant rounds of gathered data in order to reduce the amount of transmitted data. A comprehensive overview of recent machine learning methods applied in WSN can be found in [35].

However, many of the aforementioned works performing classification at the sensor network focused on the reduction of latency rather than energy efficiency. It is not clear whether that is more energy efficient than transmitting and classifying the data at the end users or not, since such classification techniques require distributed feature extraction and transmission, which may be less or more energy consuming than the transmission of measurements without classification. Furthermore, learning by examples needs to process large data sets to ensure high accuracy, whereas it is not straightforward to mathematically formulate the learned models, or to have the full control over the knowledge discovery process

Fuzzy logic techniques have been also investigated in the area of patients' care to predict and categorize patients' states [36][37]. For instance, the author in [36] leverages fuzzy techniques in the development of a decision support system that optimizes the postprandial glycemia in type 1 diabetes patients, while the authors in [37] exploit fuzzy and probabilistic computing to assess breast cancer risk. In [38], the authors discuss the design and implementation of a fuzzy logic-based warning system that exploits fuzzy logic to categorize patients' states and send timely warning messages to the healthcare service providers. In [39], a fuzzy expert system is developed to classify the patients with confined or non-confined prostate cancer showing the efficiency of the presented fuzzy system compared with other probabilistic systems.

Regarding the data compression, number of biosignal compression algorithms were proposed in the literature [40], which vary in the lossiness, computational complexity, and waveform transformation (e.g., Discrete Wavelet Transform (DWT), Autoencoders, vector quantization, discrete cosine transform, etc.) [41][42]. In [43], the authors presented lossless/near-lossless compression algorithms for multichannel biomedical signals using information theory and signal processing tools through leveraging the spatial and temporal redundancies in biomedical signals. However, the intensive computational complexity of such techniques might turn the in-network processing on battery-operated devices impractical [44][45][46]. Furthermore, non of the aforementioned works have considered the characteristics of the gathered data, or the class of the patient before compression, in order to adapt the proposed compression techniques based on the class of the data and application's requirements.

The enormous advances in smartphone capabilities have also motivated the development of smartphone apps for healthcare monitoring. Leveraging built-in sensors of

the mobile phones, smartwatchs, gyroscopic sensors, and GPS modules have enabled developing different apps for seizures detection at the smartphone. For instance, "Epdetect" application employs signal processing techniques to differentiate between normal movements and those associated with seizures [47]. When any abnormal movements are detected, this app triggers seizures detected alarm. Seizario [48] is another mobile app that uses only smartphone to detect seizures convulsions and falls exploiting accelerometer-based learning algorithms with elaborate finite-state-machines. However, such apps that relay on movements detection instead of the analysis of EEG signals are not reliable for detecting absence seizures that do not result in convulsions.

Accordingly, leveraging higher levels of autonomy and intelligence at the network edge through moving processing and classification tasks to the mobile edge node can significantly enhance energy consumption, as well as latency and response time, while satisfying the requirements of smart healthcare services.

2.3 System Model

Motivated by the edge computing paradigm, where we push the computational intelligence closer to the patient, we propose and implement s-health system, shown in Figure 2.1. This system aims for detecting patient's state exploiting feature extraction and fuzzy classification at the network edge. In particular, the proposed system is organized in a three-tier (i.e. cloud/edge/IoT devices) architecture that provides the gathered data/emergency notification to the healthcare provider. In this architecture, the EEG data is collected from a patient using an EEG headset in the IoT devices layer. Then, it is periodically transferred to a PDA, i.e., a smartphone, that represents the Edge Layer, which processes the gathered data and forwards the processed data to the far cloud, hereinafter referred to as MHC.

This study focuses on epileptic seizure detection as an application of EEG-based diagnosis.[3] Recent studies have indeed shown that the dynamic properties of EEG signals can be effectively used to differentiate between healthy subjects and diagnosed patients with epileptic disease. We remark that our study does not

[3] Although the proposed framework focuses on EEG-based application, it can be easily extended to a range of applications which are typically at a low data rate, or at higher data rates such as video streaming.

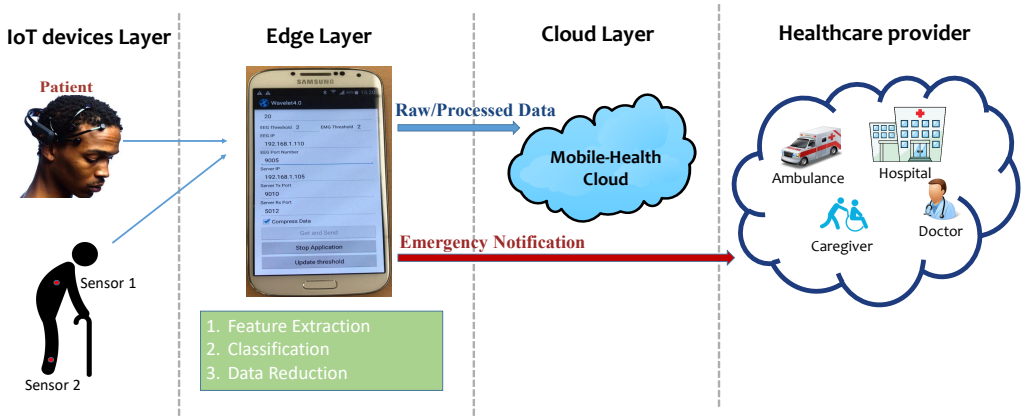


Fig. 2.1 System model under study.

only focus on monitoring people who suffer from active epilepsy, but also considers normal people who are more susceptible to seizures (i.e., high risk people). For instance, people who have surgery and became seizures free are able to stop seizures medicine. However, they may need to stay on monitoring to prevent seizures from coming back, even after becoming normal [49]. Also, seizures do happen frequently in people who have had a traumatic injury to the brain. Most seizures occur in the first several days or weeks after the brain injury, however some cases may appear months or years after the injury [50]. These kind of seizures free people we can not consider them as epileptic. However, they are more susceptible to seizures. Thus, it is of prominent importance to monitor such high risk people for seizures.

In order to conduct our study, we leverage the EEG database in [51] considering three classes of patients: seizure-free (SF), non-active (NAC), and active (AC). The first one includes seizure-free subjects (i.e., do not have seizures), the second refers to non-active patients diagnosed with epileptic disorder, however they are in non-active state, while the third class comprises patients with active epileptic seizure, as shown in Figure 2.2. Each class includes 100 single-channel EEG segments (i.e., 100 rows), and, given a sensing time frame of 23.6 s and a sampling rate of 173.6 sample/s, for each channel there are 4096 samples (i.e., columns).

With the aim to develop an energy-efficient monitoring system, we design a mechanism that enables a PDA to always select the most appropriate configuration for transmitting the patient's information, based on the patient's state. The proposed scheme is depicted in Figure 2.3.

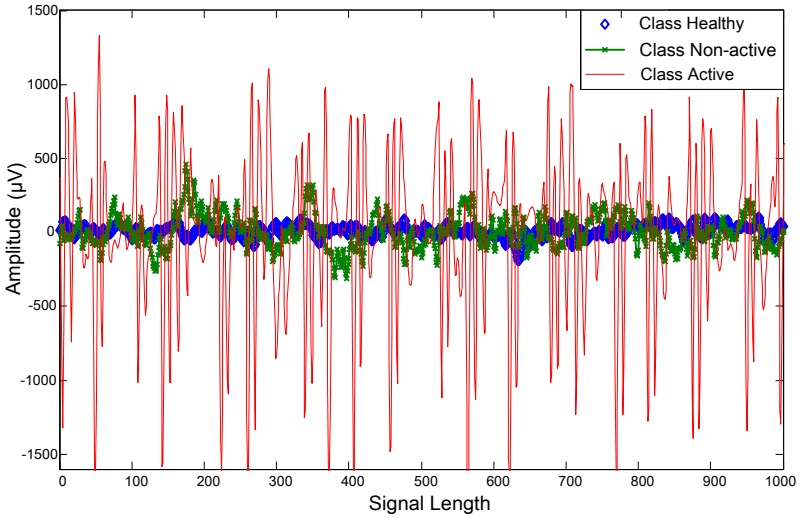


Fig. 2.2 Representation of the three classes of EEG signals in the time domain.

Starting from the collected EEG data, the PDA first derives specific values (features) that are informative, non-redundant, and pertinent to seizures detection. These features allow the classification process at the PDA (i.e., the mobile edge), as well as at the MHC when necessary, leading to an accurate interpretation of the patient's state. Based on the detected patient's state, the PDA will act as follows:

- in case of AC (i.e., emergency), it will send toward the healthcare service provider an Emergency Notification (EN) signal, along with raw EEG data to the MHC for high-intensive monitoring;
- in case of NAC, it will compress and, then, forward EEG data;
- in case of SF, it will send only EEG features (i.e., frequency-domain or time-domain features).

At the MHC, according to the received data, signal reconstruction, feature extraction, classification, or distortion evaluation can be performed, in order to accurately evaluate the state of the patient.

It is worth mentioning that for SF it is important to monitor patient's state through sending EEG features to confirm the status stability, whereas sending raw or compressed signal will not add much information to the physicians as long as the state is stable (no further analysis is needed in this case). While for NAC, it is important for the physicians to analyze the EEG signal not only the features, so they

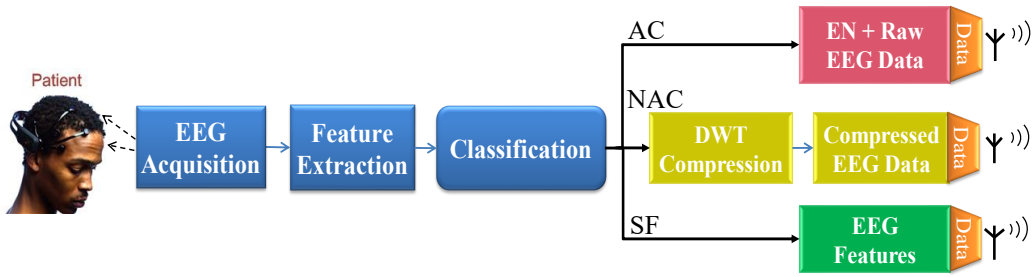


Fig. 2.3 Proposed in-network processing tasks at the edge.

can expect when seizures could happen. Thus, it is worth to send the compressed data with acceptable level of distortion to the MHC. In case of emergency (AC), it is important to send raw data without distortion for enabling sophisticated analysis of EEG signals at the cloud.

2.4 Feature Extraction

The first step in our procedure toward the design of a reliable and energy-efficient system for epileptic seizure detection and notification, consists in identifying a set of epileptic-related features through executing feature extraction on the gathered data at the PDA. To this end, we leverage two approaches: time-domain and the frequency-domain feature extraction.

2.4.1 Time-Domain Feature Extraction (TD-FE)

Our aim is to pick out the most representative time features to be used in identifying the different EEG classes. From the signal behavior shown in Figure 2.2, the three EEG classes under study exhibit different mean and variance values, in addition to different amplitude variations over time. To account for the latter, it is crucial to consider as relevant feature the waveform length, as a representation of the signal variation over time, i.e., the cumulative length of the waveform over a given time window. Moreover, it is important to consider the Auto-regression (AR) coefficients as they provide a smooth and compact representation of the signal spectrum [52][53]. We therefore select the following four statistical features:

Mean absolute value

$$\tilde{\mu}_j = \frac{\sum_{k=1}^{N} |x_j(k)|}{N} \tag{2.1}$$

Variance

$$\tilde{\sigma}_j^2 = \frac{\sum_{k=1}^{N} x_j^2(k)}{N-1} \tag{2.2}$$

Waveform length

$$WL_j = \sum_{k=1}^{N-1} |x_j(k+1) - x_j(k)| \tag{2.3}$$

Auto-regression coefficients

$$\tilde{x}_j(k) = \sum_{i=1}^{p} a_i x_j(k-i) + e_k \tag{2.4}$$

where:

- N is the considered time window expressed in number of samples, namely $N = 4096$ samples,
- $\tilde{x}_j(k)$ is the k-th sample, $k \in \{1,..,N\}$, referring to the generic patient j,
- a_i represents the auto-regression coefficient, p is the order of the auto-regression model, and e_k is the residual white noise [54].

Accordingly, for a given patient, the above four time features are representative of the patient's state over a time window of N samples.

2.4.2 Frequency-Domain Feature Extraction (FD-FE)

By transforming the gathered EEG signal into the frequency domain,[4] we observe that the different EEG classes have diverse amplitude range (see Figure 2.4) – an important characteristic that significantly facilitates discrimination between the different classes. Also, the main advantage of leveraging FD-FE is their immune to signal variations resultant from electrode placement or physical characteristics of patients [56]. Thus, by segmenting the frequency spectrum of the EEG signal into multi-subbands, each of which includes a certain number of frequency components,

[4]The EEG signal is transformed into the frequency domain using the Fast Fourier Transform (FFT) [55]. FFT is considered as a classic frequency analysis method with complexity $O(N \log N)$.

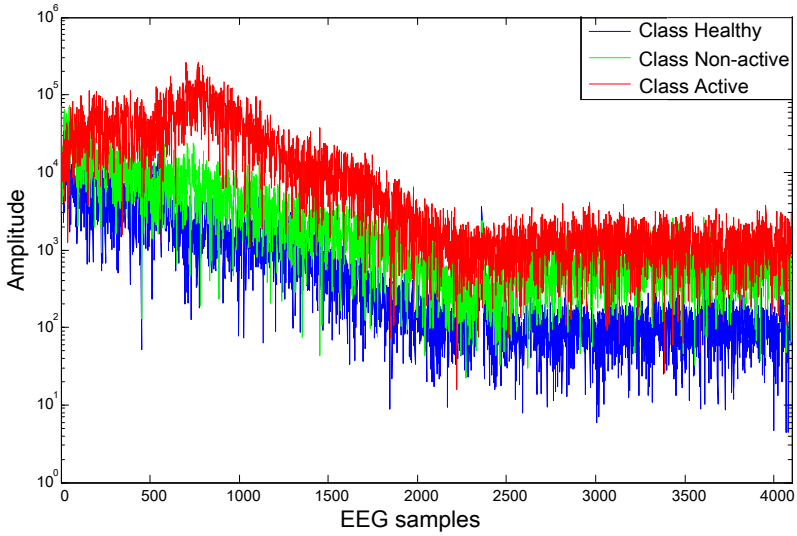


Fig. 2.4 The three classes of EEG signal after FFT.

different subsets of these sub-bands can be selected as features vector [57]. Specifically, we use the following five frequency sub-bands, named α_f, β_f, δ_f, γ_f, and θ_f, corresponding to the frequency ranges $8-12$, $12-32$, $0.2-3$, >32, and $3-8$ Hz, respectively [58]. Clearly, the more the frequency subsets that we consider, the larger the amount of data to be processed (and then transmitted), which in turn increases the energy consumption while providing higher classification accuracy.

2.5 Swift In-network Classification

The second step in our procedure toward a reliable and energy-efficient detection of epileptic seizure is classification of the patient's state. In particular, we propose a fuzzy classification mechanism to detect different EEG classes at the PDA leveraging the feature extraction techniques described above. As mentioned before, such classification, named SIC (Swift In-network Classification), has two advantages. First, it allows the PDA to select the most convenient transmission option given the detected state of the patient. Second, in case of emergency, a quick alert and notification can be initiated, which saves significant delays resulting from transmitting then classifying the data at the MHC.

In what follows, we first review some basic definitions from relational algebra as well as Formal Concept Analysis (FCA), for analyzing data and formally representing

conceptual knowledge. Then, we introduce an automated method to transform EEG signal into a fuzzy binary relation. The resultant relation is decomposed into a set of optimal concepts to build association rules for a fast, yet accurate, classification.

2.5.1 Using Knowledge Discovery in EEG Datasets

This section introduces the basic notions that will be utilized to induce a crisp relation from a fuzzy one [59], and to create a set of association rules from the obtained crisp relation [60][61]. Let $\mathscr{O}$ be the set of patients (i.e., objects) and $\mathscr{P}$ the set of features (i.e., properties). The fuzzy relation on the universe $\mathscr{U} = \mathscr{O} \times \mathscr{P}$ measures the strength of the correlation between patients and features. In order to proceed further, we recall the following formal definitions [62][63]:

Definition 1. A fuzzy binary relation, R, on the universe $\mathscr{U} = \mathscr{O} \times \mathscr{P}$, is a fuzzy set defined on $\mathscr{U}$, such that for any given pair (o, p), where $o \in \mathscr{O}, p \in \mathscr{P}$, $\mu_R(o, p)$ is the value of the membership function within R, representing the degree to which o and p are related under R.

Definition 2. Let $\alpha \in [0, 1]$. The α-cut of R, denoted by R_α, is a crisp binary relation such that, for all $(o, p) \in \mathscr{U}$, $\mu_{R_\alpha}(o, p) = 1$, if $\mu_R(o, p) \geq \alpha$. Else, $\mu_{R_\alpha}(o, p) = 0$.

Definition 3. A rectangle of R_α, denoted by $(\mathscr{A}, \mathscr{B})$, is a Cartesian product of two subsets $\mathscr{A} \subseteq \mathscr{O}$, $\mathscr{B} \subseteq \mathscr{P}$, such that $\mathscr{A} \times \mathscr{B} \subseteq R_\alpha$.

Definition 4. A rectangle $(\mathscr{A}, \mathscr{B})$ is said to be maximal under relation R_α if $\mathscr{A} \times \mathscr{B} \subseteq \hat{\mathscr{A}} \times \hat{\mathscr{B}} \subseteq R_\alpha \Rightarrow \mathscr{A} = \hat{\mathscr{A}}$ and $\mathscr{B} = \hat{\mathscr{B}}$.

Definition 5. A maximal rectangle $(\mathscr{A}, \mathscr{B})$ is said to be optimal if it maximizes the gain function. The gain function of a rectangle $(\mathscr{A}, \mathscr{B})$ is given by: $G(\mathscr{A}, \mathscr{B}) = |\mathscr{A}| \cdot |\mathscr{B}| - (|\mathscr{A}| + |\mathscr{B}|)$ where $|\cdot|$ denotes the set cardinality.

Definition 6. The coverage of R_α is defined as a set of optimal rectangles $\mathscr{V}$ under R_α such that any element $(a, b) \in R_\alpha$ is included in at least one rectangle of $\mathscr{V}$.

Examples illustrating the above definitions can be found in [59][64].

As mentioned, in our case R represents the correlation between patients and features, which can be converted into a crisp binary relation, R_α, by adjusting the

threshold α. Thus, an optimal rectangle corresponds to the maximum number of patients that share the maximum number of features can be obtained. Our aim is to obtain the minimal set of optimal rectangles covering our binary relation.

To this end, given R_α, we adopt the decomposition of a binary relation presented in [65], which is based on difunctional decomposition. Accordingly, first the Fringe Relation of a binary relation is calculated. This fringe relation is, by definition, a difunctional relation, and all its elements are isolated points. If (a,b) is an isolated point, by definition it is included in one maximal rectangle only [65]. Hence, it follows that the maximal rectangles can be easily obtained by finding such isolated points.

We then select the optimal rectangles and consider that each of them is an equivalent representation of an association rule whose head is a class label (e.g., SF, NAC, or AC). Such rules are utilized to build our classifier.

The steps we follow in order to extract the association rules from the EEG data and to classify a patient's state are exemplified in the next section.

2.5.2 Rule Extraction and Classification

For the sake of clarity, we describe the adopted procedure by referring to a toy example where the data used as training set refers to nine patients (three for each class).

Step 1: Feature extraction. Consider the patients' raw EEG samples that are available as training set. We first extract features from the collected EEG samples, using the TD-FE or FD-FE schemes presented in Section 2.4. As an example, Figure 2.5 illustrates the features obtained when the TD-FE technique is applied. These features are computed using equations (2.1)-(2.4) over the data of nine patients belonging to three classes of EEG data, namely, SF, NAC and AC.

The features are then assessed and selected. We do so by calculating the correlation of these features with the different patient classes: the features that are highly correlated with a class label, and low correlated with each other, are the most informative ones, and are thus selected.

Step 2: From feature values to fuzzy relation. In order to transform the selected features into a fuzzy binary relation, negative feature values are multiplied

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	Label
	Mean	waveLen	Variance	Auto-regression										
O1	33.954	-4	1860.9	1	-1.8722	1.0928	0.077665	-0.3711	0.20731	-0.05765	-0.26447	0.54903	-0.31912	Class Healthy
O2	37.362	41	2232.3	1	-1.3363	0.61118	0.64312	-0.29699	-0.26452	0.22298	0.034699	-0.06419	0.005384	Class Healthy
O3	42.011	28	2778.6	1	-1.5841	0.5435	0.54898	-0.39885	-0.2466	0.24142	0.016015	0.01941	-0.09515	Class Healthy
O4	53.58	-21	4622.7	1	-1.7226	0.58155	0.50927	-0.3617	-0.28726	0.38268	0.006498	-0.17789	0.086971	Class Non-active
O5	52.396	21	4615.8	1	-1.8318	0.73531	0.39856	-0.29223	-0.0961	0.16485	-0.18752	0.21037	-0.09022	Class Non-active
O6	51.404	42	3187.4	1	-1.5447	0.60515	0.41231	-0.26634	0.01629	-0.01061	-0.08779	0.23703	-0.14584	Class Non-active
O7	97.287	-193	15377	1	-2.0392	1.2561	0.26063	-0.61384	-0.0432	0.35107	-0.02676	-0.21225	0.12005	Class Active
O8	140.7	-311	38278	1	-2.2408	1.6277	0.10895	-0.64396	0.005835	0.37968	-0.15715	-0.11325	0.08825	Class Active
O9	168.15	-334	46189	1	-2.3581	1.8027	0.10798	-0.76958	0.022686	0.413	-0.08415	-0.23024	0.13719	Class Active

Fig. 2.5 Step 1: EEG time-domain features computed over the data of nine patients belonging to the three classes, namely, SF, NAC and AC.

	B1	B3	B6	B8	
	Mean	Variance	Auto-regression		
O1	0.2019269	0.0402888	0.60620181	0.48221108	Class Healthy
O2	0.2221945	0.0483297	0.33903589	0.3859118	Class Healthy
O3	0.2498424	0.0601572	0.30149221	0.51826971	Class Healthy
O4	0.3186441	0.1000823	0.32259943	0.46999662	Class Non-active
O5	0.3116027	0.0999329	0.40789371	0.3797266	Class Non-active
O6	0.3057032	0.0690078	0.33569091	0.34608488	Class Non-active
O7	0.5785727	0.3329148	0.69678815	0.79762988	Class Active
O8	0.8367529	0.8287255	0.90292339	0.83676811	Class Active
O9	1	1	1	1	Class Active

Fig. 2.6 Step 2: Transformation of EEG time-domain features into fuzzy binary relation.

by -1 and all values are normalized with respect to their maximum. The goal of the normalization is to map all selected features from Figure 2.5 onto non-dimensional values within the $[0, 1]$ range. The result is reported in Figure 2.6.

Step 3: From fuzzy to crisp. We then transform the resultant fuzzy binary relation into a crisp relation (see Definition 2), by properly setting the α parameter (see Figure 2.7).

Step 4: Finding optimal rectangles. The crisp binary relation values are then decomposed into a set of optimal rectangles (see Definition 6), using the algorithm presented in [65] and discussed in the previous section. The result of this operation in our example is shown in Figure 2.7, where we note that there is an optimal rectangle for each patient class.

Step 5: From rectangles to rules. Based on the identified rectangles, we derive a set of association rules that can be used to effectively detect the class to which patients belong (see Table 2.1). As mentioned above, in our example there are three optimal rectangles, one for each class. Given a rectangle, we create a rule whose head is given by the corresponding class label and the body is determined by the

	B1	B3	B6	B8	
	Mean	Variance	Auto-regression		
O1	0	0	1	1	Class Healthy
O2	0	0	1	1	Class Healthy
O3	0	0	1	1	Class Healthy
O4	1	0	1	1	Class Non-active
O5	1	0	1	1	Class Non-active
O6	1	0	1	1	Class Non-active
O7	1	1	1	1	Class Active
O8	1	1	1	1	Class Active
O9	1	1	1	1	Class Active

Fig. 2.7 Steps 3 and 4: Transformation of fuzzy binary relation into a crisp relation with $\alpha = 0.3$, and identification of optimal rectangles (highlighted in colors).

values taken by the selected features within the rectangle. For instance, looking at Figure 2.7, if all the patient's features take the value 1 under the crisp relation, then the patient belongs to the AC class.

However, this turned out into low classification accuracy while differentiating between the two classes SF and NAC. To enhance our approach, we therefore leverage what we called *shadow concept*: we consider not only the feature values for which the relation R_α is equal to 1, but also the negation of the features, i.e., the feature values for which the relation is equal to 0. In this case, both the features and the negation of the features of an optimal rectangle yield the condition part (body) of the rule, while the class of the patient represents its consequent part (head of the rule). Accordingly, we obtain three association rules (one for each class), as shown in Table 2.1. For completeness, in Table 2.2 we also report the association rules that we obtain applying the same procedure but using FD-FE instead of TD-FE.

Step 6: Classification. The obtained association rules are used to build a classifier at the mobile edge (the PDA). They are therefore applied to the patient's data in order to detect his/her state. Recall that, based on the detected patient's state, the PDA can select the most appropriate transmission option.

With regards to classification, while applying the above procedure to our training and validation data set, we observed that the parameter α has a strong impact on the accuracy of the classification procedure. In order to ensure high performance, we therefore perform classification in two sequential steps, each using a different value of α (namely, α_1 and α_2). At the first stage, we only differentiate between normal cases (class *SF* or *NAC*) and abnormal cases (class *AC*), using the value α_1. Then, if a normal case is detected, we move to the next stage, and, use α_2 to further

differentiate between *SF* and *NAC* patients. We remark here that the values of α_1 and α_2 are obtained during an offline training phase using exhaustive search.

Table 2.1 Association rules extracted from the relation R_α of Figure 2.7 (i.e., when TD features are used).

Rule	**Class**
If $B_1 = 0$ AND $B_3 = 0$ AND ($B_6 = 1$ OR $B_8 = 1$) Then Class Healthy	Class Healthy
else If $B_3 = 0$ Then Class Non-active	Class Non-active
else If $B_1 = 1$ AND $B_3 = 1$ AND $B_6 = 1$ AND $B_8 = 1$ Then Class Active	Class Active

Table 2.2 Association rules created using FD features.

Rule	**Class**
If (all selected features = 0) Then Class Healthy	Class Healthy
If (at least one feature = 1 AND one feature = 0) Then Class Non-active	Class Non-active
If (all selected features = 1) Then Class Active	Class Active

In nutshell, the main proposed tasks at the PDA are illustrated in Figure 2.8. In case of AC detection, our system triggers a mental disorder emergency signal at the local processing unit (i.e., PDA), as well as at the remote health monitoring server; and instead of sending the EEG features as in normal case, raw EEG data is sent. On the contrary, in case of NAC detection, the compressed EEG data is sent, while sending EEG features in case of SF. Thus, the PDA ensures fast emergency response, while saving energy, time, and memory space, by sending to the MHC compressed data for the NAC and by transmitting only the relevant features for the SF.

2.6 Data Compression Using DWT

The third and last component of our epileptic seizure detection system is the compressing and transferring the EEG data toward the MHC. In the following, we detail the adaptive mechanism we exploit for EEG data compression at the PDA.

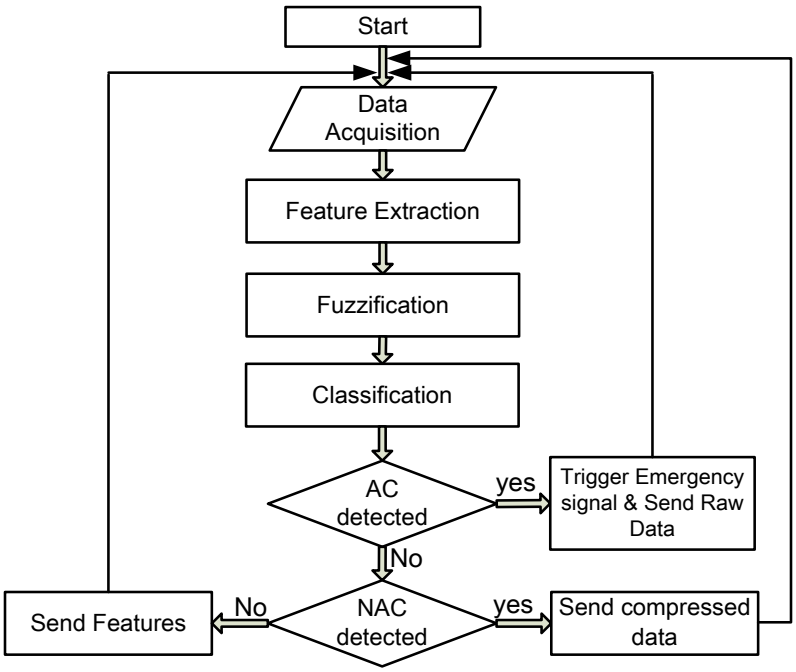


Fig. 2.8 The main proposed tasks at the PDA.

In the case of NAC, it is assumed that the PDA processes the EEG signal using the so-called threshold-based Discrete Wavelet Transform (DWT) [66], and that a Daubechies wavelet family is selected for this purpose [67]. Given signal x, we can write: $x = \Psi\alpha_w$ where Ψ is the Daubechies wavelet family basis and α_w is the vector of wavelet domain coefficients.[5] Such coefficients are then filtered using a filter with length $F = 2\varsigma$, where ς is the order of the selected wavelet family. The longer the filter length, the higher the number of output coefficients. Next, according to the threshold-based DWT, the filtered coefficients that are below a predefined threshold are zeroed [68]. It follows that the number of output samples generated from the threshold-based DWT, hence the compression ratio, can be controlled by properly setting F as well as the value of such threshold. Indeed, the compression ratio (expressed as percentage) is given by:

$$C_r = \left(1 - \frac{M}{N}\right) \times 100 \tag{2.5}$$

[5]Note that, in the case of multistage DWT, these coefficients are calculated recursively on multilevel wavelet decomposition.

where M is the number of output samples generated after the threshold-based DWT, and N is the length of the original signal. The encoding distortion caused by the compression can then be defined by the percentage Root-mean-square Difference (PRD) between the reconstructed EEG signal and the original one, as

$$D = \frac{\|x - \hat{x}\|}{\|x\|} \cdot 100, \tag{2.6}$$

where x and $\hat{x}$ are the original and the reconstructed signal, respectively.

It is important to remark here the well-known fact observed in practical system design: there is always a tradeoff between energy consumption, system complexity and encoding distortion. Our experimental results, depicted in Figure 2.9, confirm that the main parameters affecting the encoding distortion are the wavelet filter length (F) and the threshold value. The plot shows both the compression ratio and the distortion that are obtained by varying the two parameters. As mentioned, given F, the higher the threshold, the larger the number of samples that are zeroed, hence the higher the compression ratio. The reduced amount of data to be transferred clearly translates into a lower energy consumption but at the expense of an increased distortion. When we fix the threshold value, an increasing F (i.e., a higher order of the Daubechies wavelet family) leads to a larger number of output samples and a more detailed representation of the signal, which reduces distortion. Interestingly, when the threshold value is small, the compression ratio grows quite noticeably with increasing F since the generated coefficients exhibit a smaller value and are therefore zeroed when thresholding is applied (see the green curves in Figure 2.9). The price to pay for such better performance is an increased computational complexity. Accordingly, our adaptive compression technique enables the PDA to establish the preferred tradeoff by properly adjusting the encoder parameters, namely, F and the threshold value.

2.7 Automated Class-based Compression

Now, we extend the aforementioned work, leveraging the concept of applying swift classification at the edge, in order to propose an automated Class-based Compression mechanism (CbC) [22]. This mechanism will allows for decreasing the amount of transferred data, according to the class of patients. Using such an approach, the

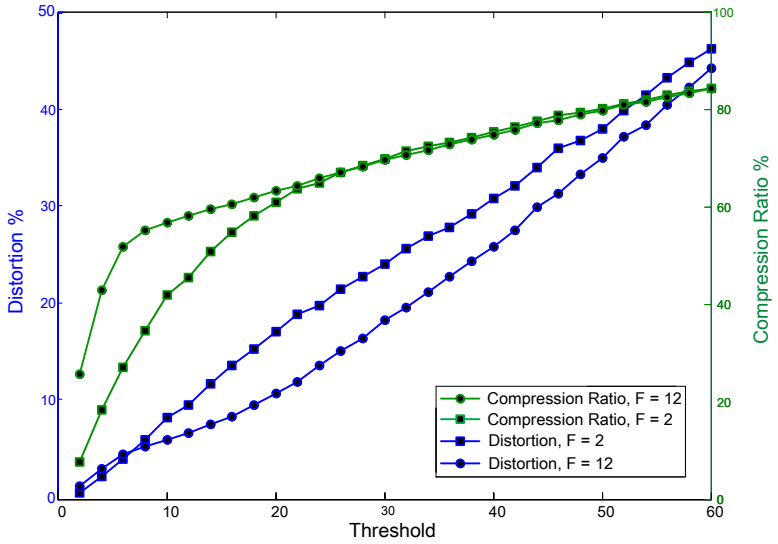


Fig. 2.9 Distortion versus compression ratio. Different values of filter length are considered.

PDA can automatically reconfigures its compression parameters based on the characteristics of the gathered data, while maintaining the application QoS requirements (i.e., signal distortion and classification accuracy). To the best of our knowledge, performing class-based data reduction at the network edge to minimize the transmission energy, while maintaining applications' QoS requirements has not been studied before.

Figure 2.10 depicts the main proposed tasks at the PDA toward implementing our CbC scheme. In particular, we propose the following tasks at the PDA:

1. Transforming EEG data into frequency domain and extracting frequency-domain features, which are informative, non-redundant, and pertinent to epileptic seizure.

2. Performing a low-complexity classification using extracted features in order to differentiate between normal/abnormal EEG signals.

3. Compressing EEG data before transmission leveraging a reconfigurable or adaptive compression threshold that is varying based on the identified class.

Accordingly, we decrease transferred data size by compressing the data, while retrieving the original data at the MHC without affecting application's QoS requirements. In what follows, we investigate the above tasks in details. Herein, we consider that

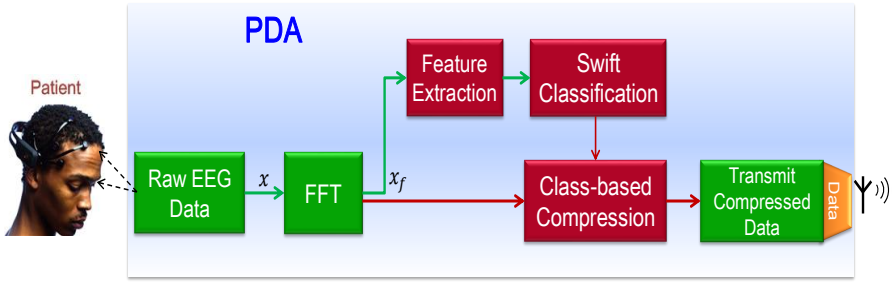


Fig. 2.10 Proposed processing tasks at the PDA toward implementing a CbC mechanism.

class SF and NAC represent the normal EEG pattern, while class AC representing the abnormal EEG pattern.

2.7.1 Feature Extraction

In our CbC scheme, we consider Frequency Features (FF). However, we leverage here the variations in the mean, median, and amplitude of normal/abnormal EEG patterns after FFT to construct our features vector (i.e., different from the previous-selected frequency features vector in section 2.4.2). Furthermore, it is crucial to consider as relevant features the Root Mean Square (RMS) to distinguish between seizures and non-seizure events, and Signal Energy (SE). RMS and SE are good signal strength estimators in different frequency bands. We therefore select the following five frequency features:

Mean absolute value

$$\mu_f = \frac{\sum_{k=1}^{N} |x_f(k)|}{N} \tag{2.7}$$

Median

$$M_f = \begin{cases} |x_f(\frac{N+1}{2})|, & \text{if } N \text{ is odd} \\ \frac{|x_f(\frac{N}{2})|+|x_f(\frac{N}{2}+1)|}{2}, & \text{if } N \text{ is even} \end{cases} \tag{2.8}$$

Peak absolute value

$$P_f = \max(|x_f|) \tag{2.9}$$

Root mean square

$$R_f = \sqrt{\frac{1}{N}\sum_{k=1}^{N}|x_f(k)|^2} \tag{2.10}$$

Signal energy

$$E_f = \sum_{k=1}^{N}|x_f(k)|^2 \tag{2.11}$$

where $|x_f|$ is the absolute value of input EEG signal x after FFT.

2.7.2 Classification

The second step in our CbC scheme is proposing a reliable classification rule for epileptic seizure detection at the edge (i.e., PDA). We leverage the extracted frequency features to perform an initial classification on normal/abnormal EEG patterns. The benefit of such classifier is that, by knowing the data class at the transmitter, we can enhance the performance of our compression technique through increasing/decreasing compression threshold without violating distortion threshold imposed by the application.

The main question here is: How can we determine a simple yet accurate classification rule using generated FF in order to distinguish between normal/abnormal EEG patterns? First, we define a classification indicator $\tilde{\lambda}$ that integrates the generated FF as follow

$$\tilde{\lambda} = \mu_f + M_f + P_f + R_f + E_f. \tag{2.12}$$

Second, we define a classification rule based on $\tilde{\lambda}$, in which $\tilde{\lambda}$ will represent the condition part of the rule, while the status of the patient S will represent its consequent part. Accordingly, we obtain through our experiments the following classification rule

$$S = \begin{cases} \text{Normal}, & \text{if } \frac{\tilde{\lambda}}{\vartheta} \leq \zeta \\ \text{Abnormal}, & \text{if } \frac{\tilde{\lambda}}{\vartheta} > \zeta \end{cases} \tag{2.13}$$

where ϑ is a scaling factor, and ζ is the classification threshold that is defined during an offline training phase based on classification indicator values for different EEG signals behavior. This classification rule is exploited to determine the state of the patient at the PDA, hence, adapting our CbC scheme.

2.7.3 Adaptive Class-based Compression

The third step in our scheme is developing an adaptive class-based compression by adjusting the transferred data size based on patient's state (i.e., class of the data). After transforming the collected EEG data into the frequency domain, the FFT returns N complex numbers (coefficients) corresponding to the N input samples. However, the generated spectrum is conjugate even (i.e., two-sided spectrum); the magnitude spectrum is symmetrical (see Figure 2.11). Leveraging such characteristics in the frequency domain, we can only transmit one-sided spectrum, hence, the output after the FFT will be $N/2$ complex coefficients. Furthermore, the coefficients that are below a predefined threshold $\tilde{\delta}$ can be discarded without much signal quality loss. Accordingly, by properly adjusting such a threshold we can control the length of the output data generated from CbC and, hence, the compression ratio of the CbC.

At the receiver side, the reconstruction and data recovery is performed using IFFT to retrieve the original signal. To quantify the difference between the original and the reconstructed signal, the signal distortion is evaluated as in (2.6).

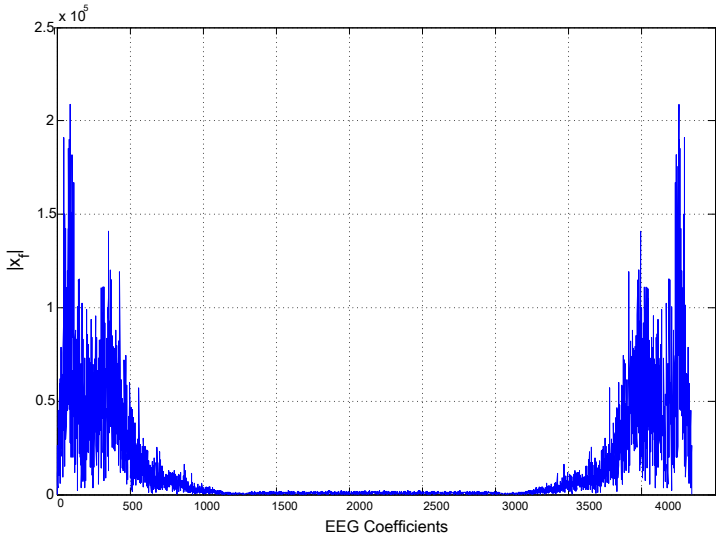


Fig. 2.11 An example of abnormal EEG signal coefficients after FFT.

The question now is: How can we define the threshold $\tilde{\delta}$? It is a fact that, in lossy compression techniques, there is always a tradeoff between increasing compression ratio and decreasing distortion. Hence, it is important to maximize compression ratio, for saving energy consumption, without violating application QoS requirement, i.e., distortion. To consummate this, we propose an Automated Seizure Detection (ASD)

algorithm. This algorithm enables the PDA to automatically update its compression threshold, hence the compression ratio, based on the class and the characteristics of the gathered data, such that it can satisfy application distortion constraint. Leveraging the extracted FF, ASD algorithm can detect normal/abnormal EEG patterns, hence update threshold $\tilde{\delta}$ as follows:

$$\tilde{\delta} = \begin{cases} \mu_f \cdot \frac{\tilde{\lambda}}{\vartheta} \cdot \upsilon, & \text{if } S \text{ is Normal} \\ \mu_f \cdot \frac{\tilde{\lambda}}{\hat{\vartheta}} \cdot \upsilon, & \text{if } S \text{ is Abnormal} \end{cases} \tag{2.14}$$

where υ is an optional tuning parameter for a user to increase/decrease compression ratio, ϑ and $\hat{\vartheta}$ are normalizing parameters for normal and abnormal EEG pattern, respectively. The main steps of the proposed ASD algorithm are illustrated in Algorithm 1.

Algorithm 1 Automated Seizure Detection (ASD)

Input:
x: Collected EEG signal.
Compute x_f.
From x_f, generate frequency features using equations (2.7)-(2.11).
Compute $\tilde{\lambda}$, as in (2.12).
if $\frac{\tilde{\lambda}}{\vartheta} \leq \zeta$ **then**
    Normal EEG pattern detected.
    Update the value of $\tilde{\delta}$ as in (2.14).
else
    Abnormal EEG pattern detected.
    Generate emergency notification signal.
    Update the value of $\tilde{\delta}$ as in (2.14).
end if
Compress and transmit x_f using obtained $\tilde{\delta}$.

We remark here that leveraging the proposed CbC scheme allows for the PDA to determine the best compression threshold based on the class of the data, while maintaining application QoS requirements. Unlike the other threshold-based techniques that neglect the class of the data and define a threshold taking the conservative approach (i.e., fixing the value of the threshold corresponding to the maximum-obtained distortion for normal EEG pattern), which decreases the obtained compression ratio for abnormal EEG pattern, or using greedy approach (i.e., fixing the value of the threshold corresponding to the maximum-obtained distortion for abnormal EEG

pattern), which results in high distortion for normal EEG pattern, as will be shown in the next section.

2.8 Performance Evaluation

In this section, we investigate the performance of our s-health system using the implemented framework shown in Figure 2.12. In the following, after presenting the experimental setup, we start by comparing the accuracy level of the classification outcome obtained at the server, when the different data reduction techniques are applied. Then, we focus on the performance of the proposed SIC scheme compared with different machine learning classifiers from the literature. After that, we compare the performance of the implemented s-health system with a mobile-health monitoring system in terms of energy saving, battery lifetime, and delay reduction. Finally, we assess the performance of the proposed CbC scheme.

2.8.1 Experimental Setup

The main components and functionality of the implemented system (see Figure 2.12) can be summarized as follow:

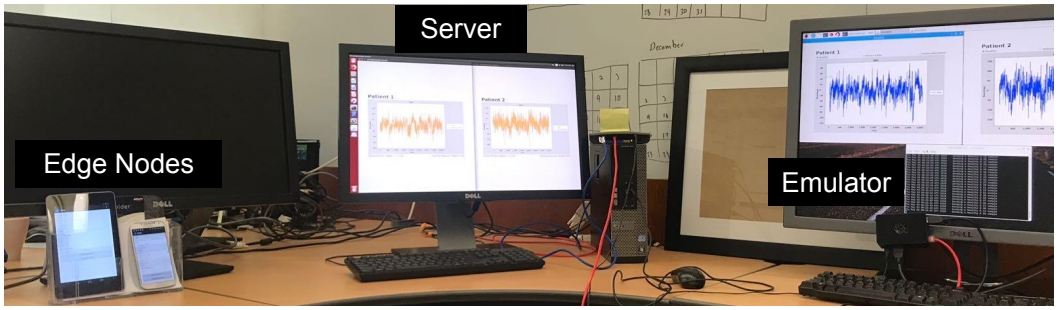


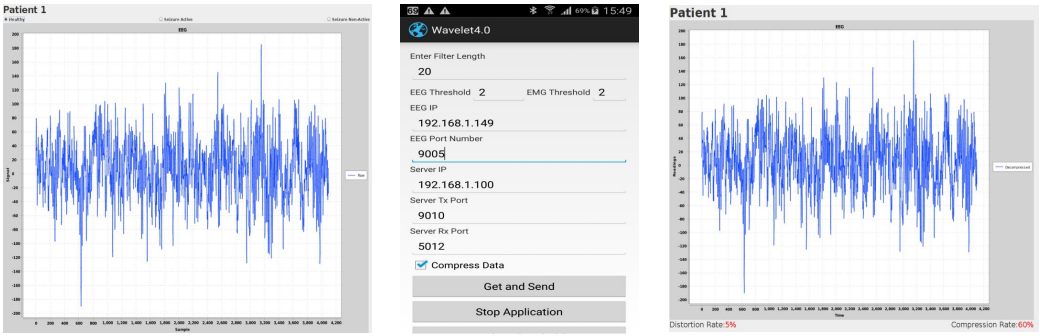
Fig. 2.12 The implemented system for EEG telemonitoring.

Data Source (i.e., Emulator). This module is responsible for generating and sending the EEG signals to the PDA. Specifically, we focus on epileptic seizure detection leveraging the EEG dataset in [51]. Thus, every 200 msec, the Emulator sends a “Medical Record" to the PDA, i.e., our application-based packet. Each record contains 4096 EEG samples and is 32 *KB* in size. The experiment ends when the server receives and acknowledges 18000 medical records.

PDA (i.e. smartphone). It is responsible for the communication with the Emulator, as well as receiving, processing, and forwarding the processed data to a health monitoring server. The communication between the PDA and the server is performed through WiFi following the IEEE 802.11 (WiFi) standard [69]. We have developed and implemented an Android application at the PDA (see Figure 2.13) that performs the following tasks:

- Swift classification using the proposed SIC algorithm in order to classify the acquired EEG signals.
- Threshold-based DWT compression, where the appropriate threshold can be adjusted based on the patient's state, desired compression ratio, application distortion threshold, and available energy budget.
- Energy-efficient transmission, where the PDA decides to send to the remote server the extracted features (i.e., time features or frequency features), compressed data, or raw data, according to the detected state of the patient, while sending an emergency notification for the patients with abnormality.

Health monitoring server. A server application is developed to receive the transmitted data from PDA, then it performs: (i) data analysis and classification if it receives raw data; (ii) classification if it receives time/frequency features; (iii) data reconstruction, distortion evaluation, and classification if it receives compressed data.



(a) (b) (b)

Fig. 2.13 Developed applications at: (a) Emulator, (b) PDA, and (c) health monitoring server.

2.8.2 Feature Extraction and Data Reduction: A Comparative Study

We first consider the FD-FE technique and the classification accuracy (CA) that it can yield. In FD-FE, the EEG signal represented in the frequency domain is segmented into multiple sub-bands, each sub-band having a number of frequency components. Different subsets of these sub-bands can be selected as features vector. Doing so we can control the amount of data corresponding to the selected features and, hence, the amount of transmitted data (see Table 2.3). Here, we assume that each sample/frequency coefficient is represented by one byte. Also, Table 2.3 illustrates the CA variations with increasing features vector length, expressed in percentage. In general, the larger the amount of transmitted data, the higher the CA, except for some cases where the added sub-bands yield a performance decrease. The reason for this behavior is that, in some cases, the added data may "confuse" the classifier rather than help. On the contrary, with increasing length of the transmitted data, the consumed energy in the transmission process always increases. Thus, an optimal tradeoff between classification accuracy and energy consumption can be established, based on the application's requirements, patient's state, and energy availability at the PDA.

Table 2.3 Selected sub-bands and corresponding amount of transmitted data as well as classification accuracy.

Frequency sub-bands					**Transmitted data**	**CA**
γ	β	α	θ	δ	**length [B]**	**[%]**
0	0	0	0	1	133	80.3
0	0	0	1	0	236	93.3
0	0	0	1	1	369	92.3
0	0	1	1	1	557	96.67
0	1	1	1	1	1501	97.78
1	0	0	0	1	2719	94.1
1	0	1	0	0	2774	87.6
1	1	1	1	0	3954	97.78
1	1	1	1	1	4087	98.89

The comparison among different data reduction techniques including TD-FE, FD-FE, downsampling (where the EEG sampling rate f_s varies), and DWT compression is presented in Table 2.4 and Table 2.5. In the case of data compression, we present

the reduction in CA compared to the case of no compression (i.e., when raw data is transmitted), which yields a CA equal to 86.67%. Note that classification based on raw data in general leads to worse performance than in the case where TD or FD features are used. The reason is again that too much redundant information may mislead the classifier rather than improve its accuracy.

We observe also that the higher the sampling rate (i.e, number of sensed EEG samples per second), the larger the amount of transmitted data and the higher the accuracy. The only remarkable exception is represented by TD-FE: in this case (i) the amount of transmitted data does not depend on the sampling rate, and (ii) we can achieve an accuracy of 95.56%, while transmitting only 13 bytes instead of 4096. When, instead, compressed data is sent by the PDA to the MHC, increasing the compression ratio leads to a slight decrease in the CA with respect to the case where raw data is transferred, while significantly reducing the amount of transmitted data. Seemingly to the effect of increasing the compression ratio, lowering f_s decreases the length of transmitted data at the expense of a reduced CA. However, it is worth noticing that in some cases a higher compression ratio, or smaller f_s, still yields satisfactory values of CA. The reason is that, in such cases, the missed data is actually redundant, thus not beneficial in terms of CA.

At last, comparing FD-FE to TD-FE and data compression, we observe that FD-FE can provide the best CA while offering significant flexibility in terms of data length: by properly selecting the subset of frequency sub-bands, the desired tradeoff between amount of transmitted data and CA can be easily obtained. However, while data compression still allows signal reconstruction at the MHC, FD-FE as well as TD-FE are irreversible: the original EEG signal cannot be reconstructed from its features, which may not be acceptable for some applications.

Table 2.4 Classification accuracy using compressed data relative to transferring raw data (which yields CA=86.67%) and downsampling.

Compressed data length [B]	**Loss in CA [%]**	**Data length [B] /f_s [sample/s]**	**CA [%]**
4096	0	4096/128	86.67
2384	3.3	2048/64	86.67
1601	1.17	1024/32	84.33
1239	0.9	819/25.6	81.33
819	2.37	682/21.3	85.67
645	1.67	585/18.3	86.3

Table 2.5 Classification accuracy of FD-FE and TD-FE.

FD-FE length [B]	**CA [%]**	**TD-FE length [B]**	**CA [%]**
4087	98.89	13	95.56
3954	97.78	13	95.56
1501	97.78	13	95.56
557	96.67	13	95.56
236	93.3	13	95.56
133	80.3	13	95.56

2.8.3 Classification Evaluation

We now focus on the performance of the proposed SIC algorithm, and illustrate the effect of the α-cut on the obtained CA at the PDA level. We evaluate the performance of our SIC algorithm when TD-FE and FD-FE are applied. Recall that for SIC-TD we use the TD-FE scheme presented in Section 2.4.1 and the association rules described in Table 2.1, while for SIC-FD, we use the feature extraction technique introduced in Section 2.4.2 and the association rules presented in Table 2.2. Also, we remark that, when working in the frequency domain, we get a much longer feature vector than in the case of TD-FE. Thus, for the sake of fairness, for FD-FE we apply the association rules only to the first 10 features.

Figure 2.14 depicts the obtained CA as the value of α varies, for the three EEG classes (i.e., SF, NAC, and AC). Herein, 270 subjects are classified by running the procedure only once (one-stage procedure). Also, we compare the CA of the proposed SIC algorithm with different machine learning classifiers, including random decision forests (RandomForest), Naive Bayes (NaiveBayes), k-Nearest Neighbors (IBk), and classification/regression trees (REPTree). Each of these classifiers is run using the default configuration in WEKA software with 5-fold cross-validation [70]. In SIC, when α is small, most of the obtained normalized features are equal to 1, while at high values of α, most of the obtained features are equal to 0. In both cases, our classifier cannot accurately differentiate between the patients' classes. In the middle region, when α ranges between 0.1 and 0.4, the value of the obtained features starts to vary between 1 and 0, which enables the SIC classifier to discriminate between different classes yielding a high accuracy. The best performance is obtained with SIC-TD and for α around 0.3, which corresponds to a CA of 82%. With SIC-FD, the best α is 0.27 leading to a CA of 64%.

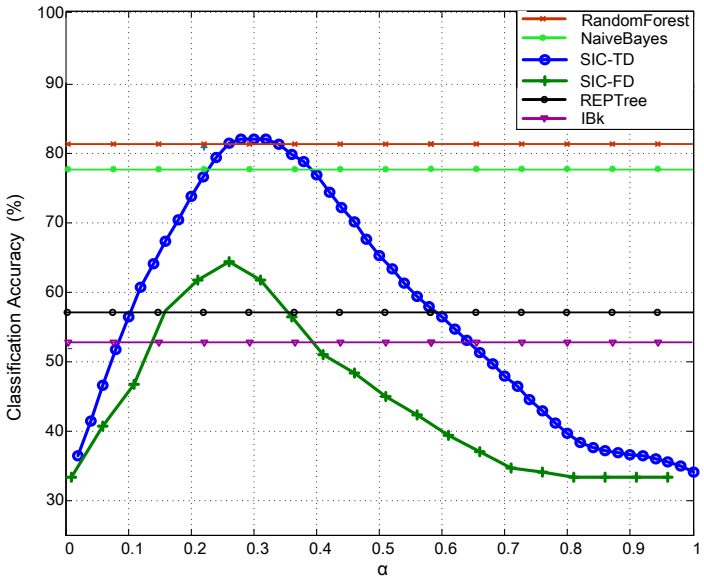


Fig. 2.14 Impact of α on the classification accuracy, for the three EEG classes.

As indicated previously, Class SF and NAC exhibit many similarities, which leads to a relatively low CA while trying to discriminate between these two classes. Thus, in order to enhance the performance of the SIC algorithm, we switch to a two-stage classification procedure using two different values of α (namely, α_1 and α_2). The results for the SIC algorithm are depicted in Figure 2.15 and Figure 2.16.

In particular, Figure 2.15 shows the results of the first classification stage as α_1 varies. In this stage, the patient's state is classified as normal and abnormal based on the gathered EEG signals. The former corresponds to class SF and NAC while the latter to class AC. The optimal value of α_1 for SIC-TD is now around 0.21 with CA equal to 98.3%, while it is 0.17 with a CA of 83% for SIC-FD.

For normal EEG patterns, we then proceed with the second stage using α_2. As shown in Figure 2.16, in this case the optimal α_2 is 0.36 with a CA 80% and 0.27 with a CA 75% for SIC-TD and SIC-FD, respectively.

In nutshell, the advantages of the proposed SIC scheme are three-fold. First, its high classification accuracy and low complexity, which turns it to be amenable for implementation in any smartphone or PDA. Second, the quick emergency notification in case of abnormality, thanks to our swift classification technique. Third, energy savings at the PDA, in case of normal EEG patterns, through transferring only the EEG features.

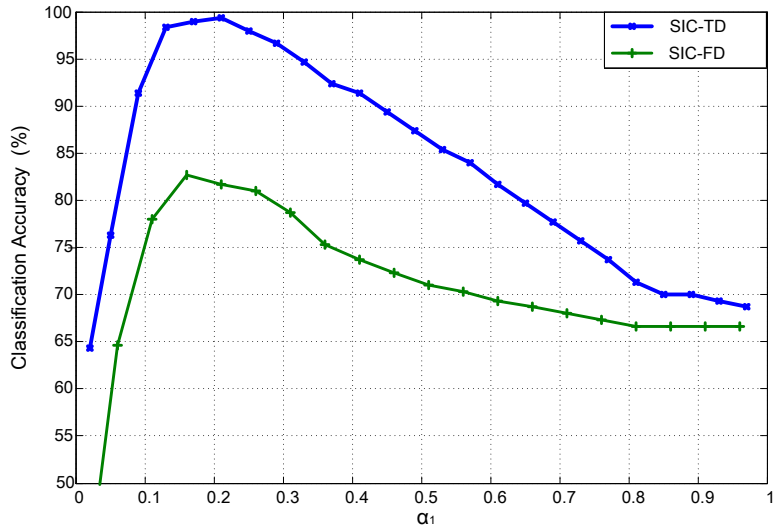


Fig. 2.15 Effect of varying α_1 on classification accuracy of normal/abnormal EEG patterns.

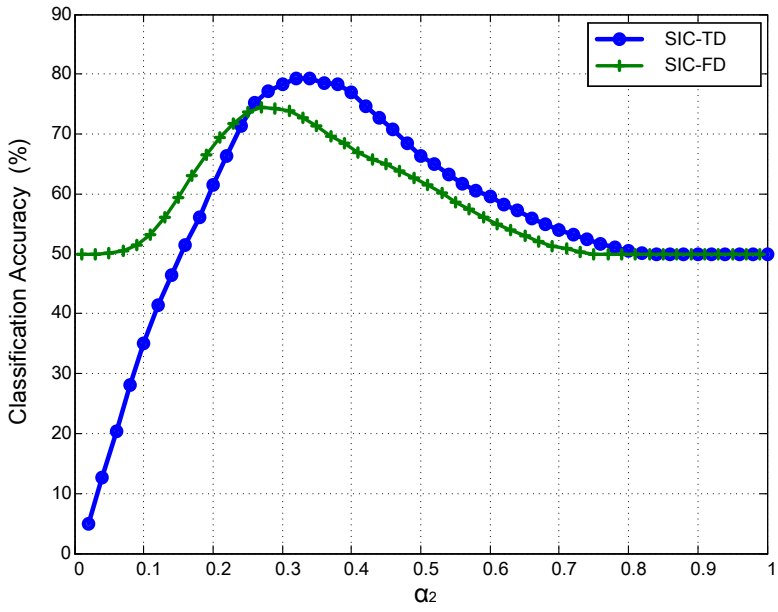


Fig. 2.16 Classification accuracy while discriminating between class SF and class NAC, as α_2 varies.

2.8.4 Energy and Delay Reduction

Here, we investigate the benefits of the proposed s-health system in terms of energy saving and delay reduction – indeed, reducing PDA's energy consumption due to continuous monitoring is one of the main objectives of this thesis. Specifically, we compare the proposed s-health system with a mobile-health monitoring system (m-health) and a remote monitoring system (RM). M-health refers to a system that acquires and transmits EEG signals from wireless sensors to the PDA, which compresses and forwards the acquired data to the server/cloud. On the contrary, RM system conveys all processing tasks to the server, while a PDA is used as a communication hub that acquires and forwards the data from wireless sensors to the server [30]. In these experiments, we analyze power usage and PDA's battery consumption using *Battery Historian* [71].

In Table 2.6, we conducted set of experiments considering a practical scenario where a PDA (i.e., smartphone) with full battery is running our monitoring application in parallel with the other default applications (e.g., Google services and Android system) until it runs out of battery. Also, it is assumed that 99% of the acquired EEG signals belong to SF, while 1% belong to AC. Table 2.6 shows the percentile of battery consumption at the PDA due to the processing and transmission of our monitoring application, running time that PDA takes until it runs out of battery, and monitoring time which is the actual time of continuous monitoring (i.e., actual time of sensing).

These experiments show the efficiency of our system and its ability to increase the monitoring time compared to RM, while decreasing battery consumption with respect to both RM and m-health. Furthermore, leveraging the proposed s-health system avoids network congestion, hence, the monitoring time is almost the same as running time, which means that the PDA is able to continuously monitor patient's state during the whole run time. On the contrary, using RM and m-health (with $C_r = 40\%$), there is a significant gap between monitoring time and running time due to the network congestion resulting from continuously sending of large volumes of data (see Table 2.6). Accordingly, the proposed s-health system has the ability to deal with a growing size of acquired data in an energy-efficient manner.

To further illustrate the advantages of the proposed s-health system in terms of energy saving, Figure 2.17 is presented. This figure assesses the performance of

Table 2.6 PDA battery consumption, running time, and monitoring time of s-health, m-health (with $C_r = 40\%$ and $C_r = 60\%$), and RM systems.

Monitoring system	**Battery consumption [%]**	**Running time [hours]**	**Monitoring time [hours]**
RM	14.71	13.5	4.5
m-health (40%)	17.62	10	8.6
m-health (60%)	16.76	11.25	10.5
s-health	11.2	10.62	10.36

s-health, in terms of PDA's battery lifetime, compared to m-health (with $C_r = 40\%$) and RM. To generate these results, we set the compression ratio for s-health in case of NAC class to 40%,[6] while considering that 10% of the acquired EEG signals belong to AC, 20% belong to NAC, and 70% belong to SF. Also, we consider the battery consumption due to our monitoring application only, while neglecting the battery consumption due to other running applications on the smartphone. Our results clearly demonstrate that s-health significantly outperforms RM and m-health with 60% and 30% extension in battery lifetime, respectively.

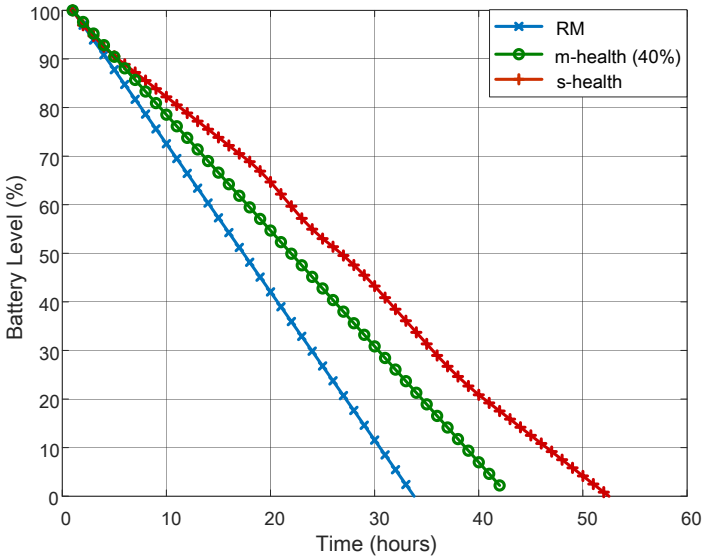


Fig. 2.17 Battery lifetime of s-health, m-health (with $C = 40\%$), and RM.

[6]The selected value for the compression ratio maintains a good tradeoff between transmission energy consumption and signal distortion (see Figure 2.9). However, other values can also be selected, according to the application requirements, patient's status, wireless channel conditions, and energy availability at the PDA.

Regarding the obtained delay reduction, Figure 2.18 shows the benefit of s-health in reducing average transmission delay. Herein, transmission delay refers to the latency experienced by the data from its receiving time at the PDA until its receiving time at the server. It is clear that reducing the amount of transmitted data using s-health has a significant impact on mitigating network congestion, hence, decreasing average transmission delay by more than 90% compared to RM.

We remark here that the obtained reduction in energy consumption and delay using s-health depends on the states of the patient, since in case of emergency (i.e., AC) the raw data should be sent to the MHC for intensive monitoring. However, the probability of AC is practically less than 1%, while seizures usually last for less than 3 minutes. After that the patient can recover his/her normal state for a long time before seizures coming back. Moreover, people who are more likely to have seizures and epilepsy (e.g., babies with abnormal areas in the brain, people with traumatic injury or serious brain injury, etc.) are usually put under monitoring after getting free of seizures for one to two years. In such cases, the proposed s-health system is recommended as an efficient solution for continuous monitoring without limiting patients' daily activities.

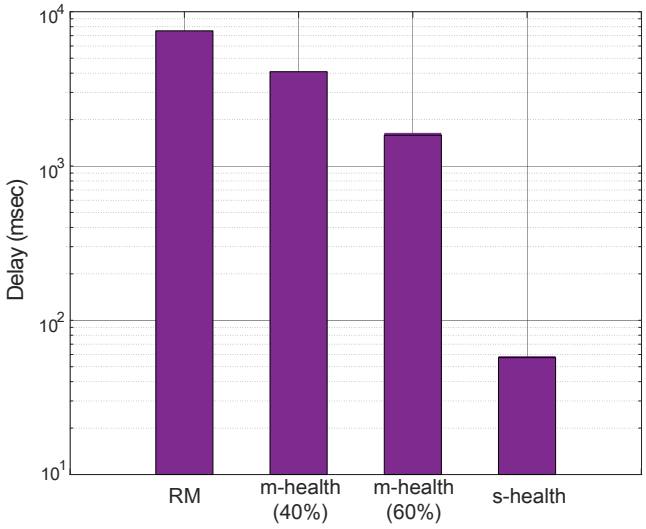


Fig. 2.18 Transmission delay of the s-health, m-health (with $C = 40\%$ and $C = 60\%$), and RM.

2.8.5 Class-based Compression Scheme

First, We assess the performance of our classification rule, presented in section 2.7.2, in differentiating between normal/abnormal EEG pattern. Figure 2.19 depicts the obtained CA using our classifier with changing ζ while considering 270 EEG records (180 for normal EEG signals, and 90 for abnormal EEG signals). At low ζ, the classifier tends to consider most of the normal EEG signals as an abnormal signals, which results in maintaining low CA. However, by adjusting the value of ζ, the classifier obtains high CA. The results show that the proposed classifier could achieve 98.3% CA with the optimal-obtained ζ, which is around 0.65. Accordingly, we could efficiently detect the emergency case (i.e., abnormal EEG signals) with very high accuracy using the proposed edge-based classifier.

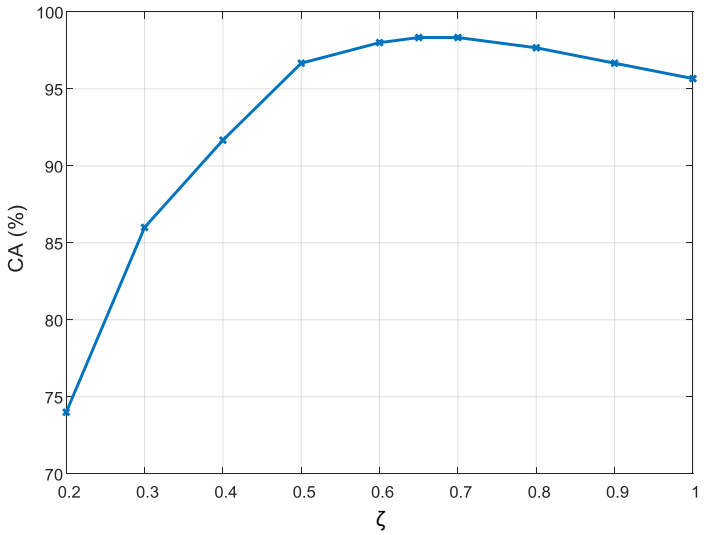


Fig. 2.19 Effect of varying ζ on obtained classification accuracy.

Next, we assess the performance of the proposed CbC technique compared to threshold-based DWT [67]. In general, by increasing the compression ratio, the distortion increases for both CbC and DWT. However, in CbC, at the same compression ratio, we could maintain less distortion than DWT (see Figure 2.20). It is worth mentioning that through varying the Daubechies families, or decomposition levels of the DWT, it can maintain less distortion, however, it comes at the expense of increasing the computational complexity, which may not be acceptable for such battery-operated devices [72].

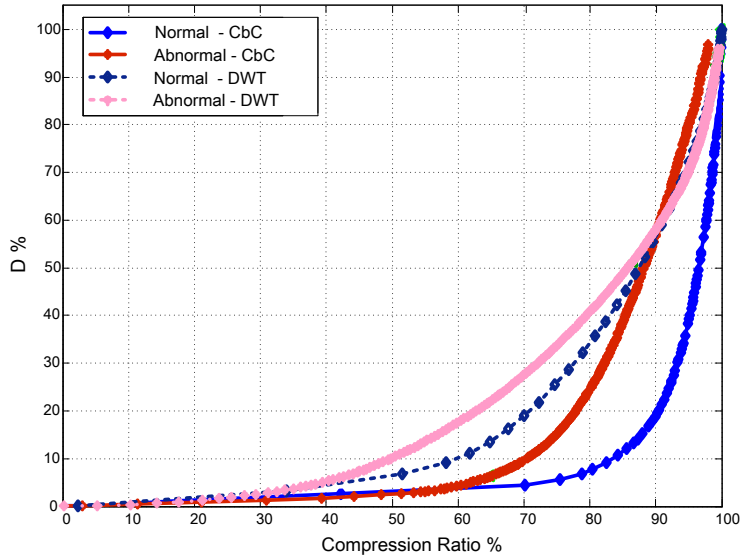


Fig. 2.20 Distortion variation with compression ratio for proposed CbC technique and DWT.

Furthermore, we remark here that for the same compression ratio, the value of the distortion varies based on the class of the data. Thus, by knowing the class of the data at the PDA, the compression ratio can be increased, while maintaining the required distortion threshold. It is clearly illustrated in Figure 2.21. As mentioned, in our CbC technique we control the transmitted data length by changing the threshold $\tilde{\delta}$: as $\tilde{\delta}$ increases, C_r increases, at the expense of increasing the distortion. As shown in Figure 2.21, at the same $\tilde{\delta}$, the distortion D and C_r vary according to the EEG class. Hence, to obtain the best C_r that satisfies application distortion threshold, the PDA should properly adjust $\tilde{\delta}$ based on the detected EEG class. Thus, it is important to have an initial-swift classifier at the PDA to obtain the proper compression threshold based on the class of the data.

Finally, Figure 2.22 illustrates the main advantage of the proposed ASD scheme compared to fixed threshold compression scheme, and assesses the ability of our scheme to adapt to varying EEG records. This figure depicts the average obtained distortion and compression ratio for each EEG class (i.e., class SF, NAC, and AC, respectively) over the time. Also, it is assumed that there is a constraint on the maximum obtained distortion, i.e., $D \leq 7\%$. We compare the ASD algorithm with two fixed threshold schemes: Conservative and Greedy compression schemes. In Conservative scheme, we consider that the threshold $\tilde{\delta}$ is fixed and identified using normal EEG class, such that the obtained distortion at normal EEG class is below the

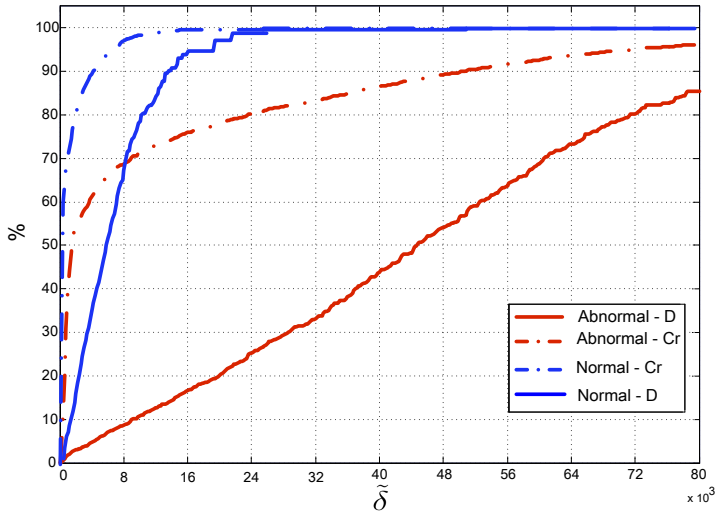


Fig. 2.21 Effect of varying the threshold $\tilde{\delta}$ on compression ratio and distortion for different EEG classes.

predefined distortion constraint. In Greedy scheme, $\tilde{\delta}$ is fixed such that the obtained distortion at abnormal EEG class is below the predefined distortion constraint. On the contrary, ASD algorithm obtains compression threshold $\tilde{\delta}$ taking into consideration the class and characteristics of the compressed data, unlike the other algorithms that consider fixed $\tilde{\delta}$ over the time. Thus, with changing the class of the collected data, ASD algorithm can adapt transmitted data size without violating given distortion constraint. On the contrary, fixing $\tilde{\delta}$ at low value (as in Conservative approach) maintains distortion constraint at the expense of obtaining very low compression ratio for abnormal class. While fixing $\tilde{\delta}$ at high value (as in Greedy approach) achieves high compression ratio at the expense of violating distortion constraint for normal class (see Figure 2.22).

2.9 Summary

Motivated by the advantages of edge computing in pushing data processing toward the data sources, we investigate wireless EEG telemonitoring system and present a full-fledged framework for seizures detection and notification. In particular, we develop and implement an accurate, lightweight classification mechanism that, leveraging some time-domain features extracted from the collected EEG data, enables a reliable seizures detection at the network edge with precise classification accuracy

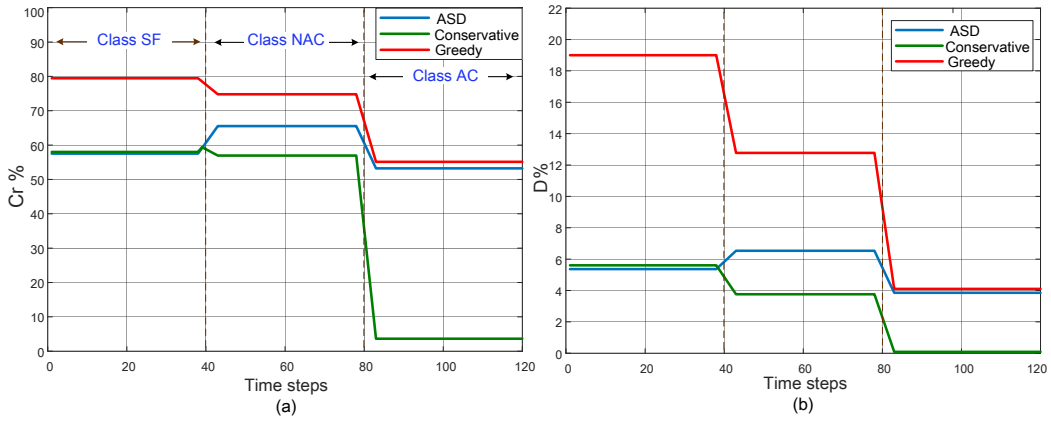


Fig. 2.22 Temporal evolution of the system performance, (a) compression ratio and (b) distortion, with varying EEG records.

and low computational complexity. In addition to that, we propose a selective data transmission scheme, which opts for the most appropriate way for data transmission depending on the detected patient's state. Finally, we propose a class-based data compression scheme that is used in epileptic monitoring systems.

Our experimental results assess the performance of the proposed system in terms of data reduction, classification accuracy, battery lifetime, and transmission delay. Specifically, the results show that:

(i) Time-domain feature extraction exhibits high classification accuracy, while minimizing the amount of transmitted data and energy consumption at the PDA. The amount of transmitted data could be decreased from 4096 to 13 samples per patient, while obtaining high seizures detection performance with a classification accuracy above 98%.

(ii) Frequency-domain feature extraction provides high flexibility, yielding the best tradeoff between classification accuracy and energy consumption.

(iii) Adaptive data compression is a very valuable option whenever more data is needed at the MHC for further processing: compared to transmitting raw data, it reduces energy consumption and, unlike feature extraction techniques, it enables signal reconstruction at the MHC.

In addition to that, we demonstrate the effectiveness of our system and its ability to outperform conventional remote monitoring systems that ignore data processing at

the edge by: (i) achieving 98.3% classification accuracy for seizures detection, (ii) extending battery lifetime by 60%, and (iii) decreasing average transmission delay by 90%.

Regarding the proposed CbC scheme, it is shown that the PDA can automatically reconfigure its compression ratio based on the class of the gathered data through adjusting its threshold, hence, save a significant amount of transmitted data while maintaining distortion constraint. Moreover, our results demonstrate that the proposed scheme outperforms the state-of-the-art compression schemes, with the advantage of reconstructing the signal at the receiver side with minimum distortion.

Chapter 3

Data-specific Transceiver Design for Healthcare Applications

3.1 Overview

Given the aforementioned requirements and challenges of e-health applications (i.e., discussed in section 1.2), designing efficient transceivers is important for s-health systems to decrease latency and energy consumption through utilizing efficient data reduction techniques. Different modules of physical layer have to be optimized in order to be flexibly configured according to the technical requirements of each application. Thus, we propose in this chapter an efficient data-specific transceiver design that leverages the inherent characteristics of the generated data at the physical layer to reduce transferred data size without adding significant overheads.[1] Our goal is to adaptively decrease the amount of data that needs to be transmitted in order to efficiently communicate with the MHC, while maintaining the required application QoS requirements.

We argue that devoting transceiver design to be specific for a certain type of data (e.g., EEG) is perfectly in consistence with IoMT devices that mostly acquire one type of data efficiently (e.g., using Emotiv headset, or QardioCore wireless ECG monitor). Hence, leveraging the characteristics of such data at the physical layer will have positive effects on the costs as well as on the energy consumption of the radio frequency (RF) transceiver. In particular, this chapter focuses on EEG signal, which

[1] This work has been published in [73].

is the main source of information on brain electrical activities that play an important role in the diagnosis of several brain disorders [24], and have a primary role in Brain Computer Interface (BCI) applications [74][75]. Thus, our main contributions can be summarized as follows:

1. Design an efficient EEG-based transceiver that leverages the characteristics of the EEG signals at the physical layer in order to provide an efficient transmission, while maintaining application QoS. Leveraging the existing orthogonal frequency division multiplexing (OFDM) transceiver's components, the proposed design performs the data compression task as part of the physical layer, hence leading to an efficient compression scheme with no significant overhead.

2. Decompose generated data into multiple streams to further increase compression ratio through discovering the dependency between different streams, and applying different compression thresholds for each stream.

3. Evaluate the proposed design through simulations discussing the tradeoff between transmitted data length and signal distortion. Our results show the gain provided by our solution, and its ability to obtain high lossless and lossy compression ratios.

3.2 Related Work

In the era of IoMT and Industry 4.0, the rapid advances in data volumes, cloud storage, edge computing, and ubiquitous network connectivity have enabled gathering, storing, and analysis of large volumes of operational data that were previously impossible. A key aspect in achieving the anticipated goals of Industry 4.0 is data processing. Given the expected volumes of gathered data, the provision of valuable services is challenging without adequate information processing and management. This is also manifested through the leading International Patent Classification (IPC) analysis, where G06F19/00 category (i.e., digital computing or data processing equipment or methods, specially adapted for specific applications) is ranked fifth in the top 20 technical IPC categories, with 130 patents between January 2006 and December 2015 [76]. Furthermore, reducing the amount of transmitted data is essential for battery-operated IoMT devices in order to save transmission energy. Some promising approaches in this context are: (i) performing local in-network

processing and compression on the gathered data before transmission, including compressive sensing (CS) [77][78][44], (ii) leveraging deep learning as a powerful tool for machine learning and health informatics to generate optimized high-level features and semantic interpretation from the gathered data [79].

In CS, sampling is performed by multiplying the original signal with a linear embedding matrix, hence, the high-dimensional data vector is projected into a low-dimensional subspace. Although, CS has shown great promise for providing high compression ratio, the construction of CS-based hardware is challenging [80]. In particular, the signal reconstruction requires in general high computational cost, which limits the use of CS in strictly real-time applications [81]. As an example, the Orthogonal Matching Pursuit (OMP) technique for signal reconstruction involves heavy matrix computation (see, e.g., [82] for details). Furthermore, there is a tradeoff between hardware energy-efficiency and signal recovery accuracy. For instance, if a data-driven optimization method (e.g., NuMax [83]) is used for improving signal recovery accuracy, more hardware power consumption is expected. On the contrary, if non-data-driven random Boolean embedding is applied for enhancing hardware energy-efficiency, high recovery accuracy cannot be guaranteed [84].

Relevant to our work are also the compression techniques that have been proposed in the literature specifically for e-health applications and that are summarized in Table 3.1. Such techniques differ in computational complexity, lossy and lossless characteristics, as well as the used waveform transformation (e.g., Fourier or wavelet transforms, vector quantization, or discrete cosine transform).

In short, most of the existing work on compression is applied at the higher layers, while ignoring lower layers features (e.g., characteristics of wireless channels, signal-to-interference-plus-noise ratio (SINR), and bit/symbol error rate). Also, consequent computational complexity might turn implementing such schemes on battery-operated devices costly.

On the other hand, designing application-specific transceivers has recently gained interest. There are some efforts on enhancing future transceivers architecture to cope with long-range IoT communication and multi-standard RF transceivers, while providing high degree of scalability, flexibility, and reusability [1][92]. For instance, SRT Marine Systems granted reversible radio transceiver US patent 9473197 and European patent EP2951930 for its reversible time domain duplex (RTDD) transceiver technology. This technology enables a specific single RF architecture to be used

Table 3.1 Summary of compression techniques for e-health applications

Application	Collected Data	Description	Advantages	Limitations
Health monitoring [85]	respiratory rate and ECG	Design an online signal compression algorithm using online codebook	Achieve high compression ratios	Large use of memory space
Compression of biometric signals [86]	ECG	Develop lossy compression algorithm using denoising autoencoders for biometric signals compression	Computationally lightweight	Limited to signals with a certain degree of periodicity
EEG remote monitoring [87]	EEG signals	Neural network-based compression algorithms proposed for near-lossless EEG signals compression using diverse neural network and linear predictors	Reduced demand for memory space and network bandwidth	Using neural networks for prediction requires training and adding computational overhead
ECG remote monitoring [88]	ECG	DCT based compression algorithm is proposed leveraging DCT comparison between different ECG peaks	Efficient ECG compression	Processing of the DCT coefficients is energy prohibitive
EEG Compression and Seizure detection [89]	EEG	Using DWT and dynamic reference lists to transmit decorrelated sub-band coefficients, while opting Set partitioning in hierarchical trees (SPIHT) as a source coder	Compressing EEG channels in one dimension (1-D), while also detecting seizure activities	Overhead added for decoding the data at the receivers side
Mental disorders monitoring [90]	EEG	Lightweight 1.5-D multi-channel EEG compression proposed using 1-D DWT and 2-D No List SPIHT arranging method exploiting correlations of wavelet coefficients	Enhance compression ratios with lower distortion while reducing power consumption	Limited performance compared to 2-D SPIHT algorithm in the near-lossless compression
Patient remote monitoring [91]	EMG and EEG	Lossy compression scheme for multidimensional biomedical signals is proposed using codebook-excited linear prediction approach and signals modeling as filtered noise	Leveraging intra- and inter-channel redundancy in time and space	Complexity and processing delay increase linearly with number of considered channels

simultaneously for both receiving and transmitting through electronic reversing of the RF chain between receive and transmit. This could be maintained through a complex combination of intelligent selection of intermediate frequencies and ultra-fast switching. Also, SRT Marine Technology Ltd of Bristol recently obtained a patent for a radio transceiver which enables multiple transceivers to share a single antenna in order to reduce costs and installation effort (GB patent no. 2460012). In the field of fiber-optic communications, [93] provides a transceiver for: (a) transmitting combined data and time code information over a fiber-optic communication means, and (b) receiving combined data and time code information from the fiber-optic communication means. In [94], the inventors present a transmitter that can receive digital audio and program information input from a plurality of sources and simultaneously broadcast a plurality of digital audio and program information signals over a limited range, hence, a user can enjoy listening to the digital audio, while enabling portable reception of the service within a localized setting.

In this context, one of the key components for the proposed transceiver design is reducing transmitted data size without adding additional hardware, while retaining the ability to reconstruct the original data from the received sequence. Also, equipped with accelerated FFT blocks, most of the current and future transceivers architecture [95][96] have the opportunity to leverage such a concept at the physical layer, saving significant amount of energy consumption and computational overheads. Our results can also be extended to have a significant practical impact in many big data domains, since most video, audio, and medical images are compressible or sparse in nature, hence, they can benefit from such compression scheme through leveraging the acquired data characteristics in the physical layer of the wireless transceiver. However, new challenges, such as the effect of quantization and modulation, have to be tackled at the physical layer to achieve such adaptive compression.

3.3 System Model

This chapter proposes an efficient transceiver design that relies on OFDM technology while obtaining an adaptive compression method to control the size of the transmitted data. OFDM is a well designed technology for high-rate wireless communication. However, the performance of such systems is generally limited by the available energy budget. Thus, we can save in energy consumption through compressing the

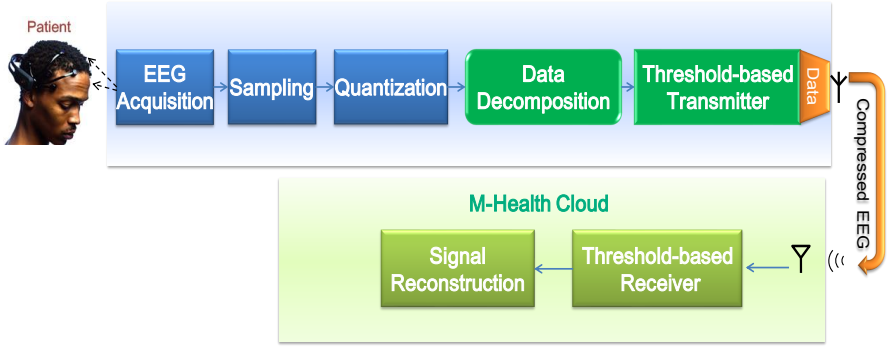


Fig. 3.1 System model under study.

data before transmission, while retrieving the original data at the receiver with zero or low distortion depending on the applied compression and application requirements.

The proposed system, shown in Figure 3.1, represents the end-to-end system from the EEG data acquisition to the MHC. The EEG data acquisition is performed using an EEG Headset [97]; the signal is then sampled, quantized, and transmitted to the MHC through our proposed transceiver. The main considered modules in our system are described below.

Sampling: Let the original continuous-time electroencephalography (EEG) waveform $s(t)$ have a duration of T seconds. The waveform is sampled at a constant interval of T_s seconds to yield N_s discrete-time consecutive samples. The sampling frequency is then given as $F_s = \frac{1}{T_s}$ Hz. Hence,

$$s_n = s(t)\ \delta\left(t - nT_s\right) = s(nT_s), \tag{3.1}$$

for $n \in \{0, 1, \cdots, N_\mathrm{s} - 1\}$, where $\delta(\cdot)$ is the Dirac delta function. Our adopted notations are reported in Table 3.2.

Quantization: The continuous amplitude of each sampled signal is quantized using an L-bit analog-to-digital converter (ADC) to one of 2^L levels, yielding the quantized signal $\tilde{x}_n$ at time index n. Each signal $\tilde{x}_n$ holds a signed integer value in the range $\{-2^{L-1}, \cdots, 2^{L-1} - 1\}$. We can express the quantized signal in vector form as

$$\underset{N_\mathrm{s}\times 1}{\tilde{x}} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{N_\mathrm{s}-1} \end{bmatrix}^\mathrm{T}. \tag{3.2}$$

Table 3.2 Summary of the used Notations

Notation	Definition
T	EEG waveform duration
N_s	Number of samples
$F_s = N_s/T$ Hz	Sampling Frequency
T_s	Inter-sample duration
L	Number of bits per sample
M	Number of symbols per sample
$K = L/M$	Number of bits per symbol

Data decomposition: The collected EEG quantized samples are written as a sequence of symbols that depend on the adopted modulation. Such symbols are then divided into multiple streams, which are processed using Formal Concept Analysis in order to discover the correlation existing between the different streams. The streams that are found to be independent of each other, are compressed and transmitted, while others are discarded. Additionally, each stream is compressed separately, using a compression threshold that suits its characteristics, thus further increasing the compression ratio.

Threshold-based transceiver: To comply with the current and future trends, we start from the typical OFDM transceiver architecture [98] (depicted in Figure 3.2-(a)), and add two simple blocks: the Threshold-based Compression (TBC) and the FFT Vector Reconstruction (see Figure 3.2). OFDM-based waveforms family is the foundation for the current LTE and WiFi systems, and is also recommended to meet the evolving requirements of 5G due to higher spectral efficiency, asynchronous multiplexing, and lower complexity. Nevertheless, our solution can be applied to other transceiver architectures such as the one adopted by the IEEE802.11ah [1] and the IEEE802.15.4g standards [96].

In the proposed transceiver architecture, it is assumed that the available quantized samples are encoded into an unsigned binary sequence via the L-bit ADC. We first turn the generated signed-integer samples into non-negative integers by a simple shift, i.e.,

$$x_n = \tilde{x}_n + 2^{L-1}. \tag{3.3}$$

Then the symbols are forwarded to the modulation and IFFT modules. Details on the TBC and FFT vector reconstruction blocks are discussed in the following sections.

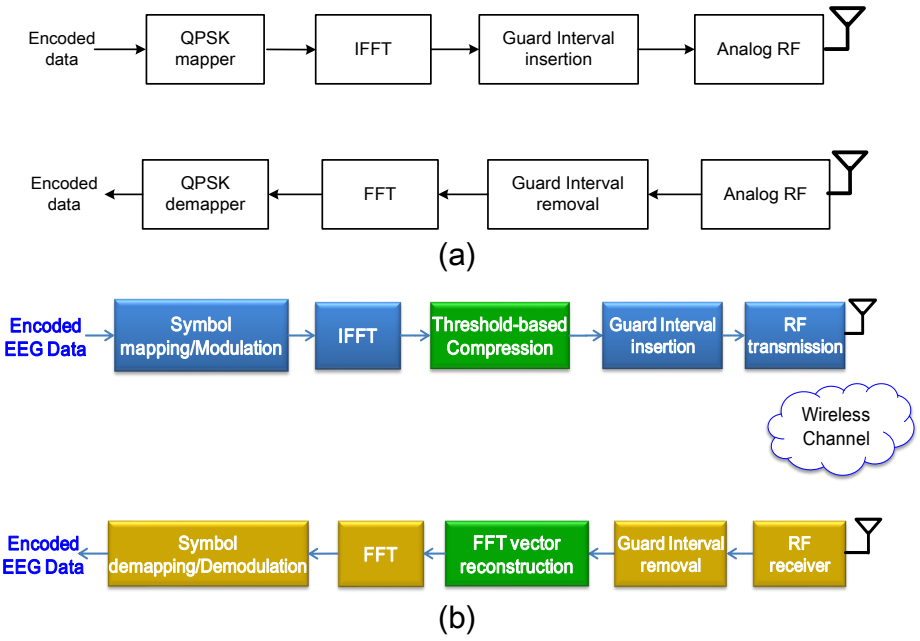


Fig. 3.2 Block diagram of (a) the basic transceiver architecture for IEEE802.11ah systems [1], (b) the adopted EEG transceiver.

3.4 Data Decomposition and Knowledge Discovery

In this section, we leverage the physical layer's characteristics to decompose quantized EEG samples into multiple streams of symbols, such that the dependency between different streams can be reduced, hence, increase compressibility.

3.4.1 Data Decomposition

First, the EEG signal x_n is decomposed into multiple streams of symbols x^m, for $m \in \{1,2,\cdots,M\}$. Let the binary encoded sequence of x_n be denoted as $b \in \mathbb{F}_{2^L}$, with $\mathbb{F}_p$ being the Galois Field of order p. Hence, b is a sequence of L bits on the form

$$b = \overbrace{\left[\underbrace{\mathrm{b}_{\mathrm{K}-1}^{(\mathrm{M}-1)}\cdots\mathrm{b}_{0}^{(\mathrm{M}-1)}}_{b^{(M-1)}} \quad \cdots \quad \underbrace{\mathrm{b}_{\mathrm{K}-1}^{(0)}\cdots\mathrm{b}_{0}^{(0)}}_{b^{(0)}}\right]}^{L \text{ bits}} \tag{3.4}$$

where $b^{(m)}$ is the group of K bits composing the m-th symbol, with $m \in \{1,\cdots,M\}$, where K and M depend on the modulation type. More specifically, M is the number of symbols per sample, which depends, not only on the modulation order O, but also

on the number of bits per sample L, as follows:

$$M = \frac{L}{\log_2(O)} = \frac{L}{K}. \tag{3.5}$$

Then, x_n can be rewritten as:

$$x_n = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} 2^{mK+k} b_k^{(m)} = \sum_{m=0}^{M-1} 2^{mK} x^m, \tag{3.6}$$

with

$$x^m = \sum_{k=0}^{K-1} 2^k b_k^{(m)}. \tag{3.7}$$

In conclusion, the bit stream block $b^{(m)}$ is simply the binary representation of x^m, which implies $x^m \in \{0,1,\cdots,2^K-1\}$. For example, using 16-QAM modulation with $L = 12$ bits, each EEG sample will be represented in three data symbols, where $M = L/4 = 3$.

3.4.2 Knowledge Discovery

Leveraging the created symbol streams, the compression ratio can be further increased by discovering the correlation between different streams. In a nutshell, using FCA for knowledge discovery [64], we can define the minimal-representative streams, so as to reduce the number of transmitted data streams, without losing knowledge.

We start by introducing the basic notions used to induce a binary relation between the generated streams. Let $\mathscr{O}$ be the set of streams (i.e., objects), $\mathscr{A}$ the set of symbols' values (i.e., attributes), and I the binary relation on the universe $\mathscr{U} = \mathscr{O} \times \mathscr{A}$ that defines which objects have which attributes. In order to transform our streams into formal context of $(\mathscr{O}, \mathscr{A}, I)$, we consider the attributes a_v of each symbol s to be all the possible values it may take, depending on the employed modulation, for $v \in \{0,1,\cdots,2^K-1\}$, and $a_v \in \{0,1\}$. Thus, the vector of attributes $\mathscr{A}$ for each stream x^m is defined as

$$\mathscr{A}(\tilde{x}^m) = \left[\underbrace{a_0 \cdots a_{2^K-1}}_{s_1} \quad \cdots \quad \underbrace{a_0 \cdots a_{2^K-1}}_{s_{N_s}} \right], \tag{3.8}$$

	S1				S2				S3				S4				· · · · ·	S17				S18				S19				S20			
	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3		0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
O1	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	· · · · ·	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0
O2	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	· · · · ·	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1
O3	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	· · · · ·	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	0
O4	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	· · · · ·	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0
O5	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	· · · · ·	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1
O6	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	· · · · ·	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0

Fig. 3.3 Steps 1 and 2: Transformation of generated streams into a binary relation, and identification of the concepts (highlighted in colors).

where $\mathscr{A}$ represents the possible values of each symbol.

Our aim now is to determine the dependency between various streams through defining the minimal set of formal concepts[2] covering our relation. Herein, we refer to the implications as the minimal set of rules, by which we can infer some attributes from others. We can derive formal concepts from our formal context using the derivation operators or difunctional decomposition[3] [65][99]. Once the formal concepts are derived, implications can be identified, hence transmitting only the minimal-representative number of streams. For the sake of clarity, we describe the adopted procedure by referring to a toy example, where a data length of 20 symbols with QPSK modulation is considered.

Step 1: Generation of formal context. Consider the generated streams of symbols. We consider each stream as an object with attributes corresponding to its symbols' values. As an example, Figure 3.3 illustrates the formal context of 6 streams with 20 symbols.

Step 2: Identifying formal concepts. The generated binary relation are then decomposed into a set of concepts, using the algorithm presented in [65]. However, in order to well identify the dependency between different streams, we leverage what we called *shadow concept*, presented in section 2.5.2. In this case, both the attributes and the negation of the attributes form the identified concept.

[2] (O,A) is a formal concept if A is the set of all attributes shared by the objects O, and in the same time O is the set of all objects that have all attributes in A.

[3] Difunctional decomposition enables obtaining the isolated points of a binary relation through calculating the Fringe Relation. This fringe relation is, by definition, a difunctional relation, and all its elements are isolated points. Thus, the formal concepts can be easily obtained by finding such isolated points, since if (a,b) is an isolated point, by definition it is included in one concept only [65].

Step 3: From concepts to implications. Based on the identified concepts, we derive the implications that can be used to effectively eliminate the streams that can be retrieved at the receiver using their implications with other received streams. For instance, looking at Figure 3.3, we can easily identify from the obtained concept that $O_2 \rightarrow O_1$, where $\rightarrow$ stands for the implications, since $O_2 = O_1 + 2$, for $O_2, O_1 \in [0, \cdots, 3]$.

Step 4: Elimination. For each obtained concept, we transmit only one stream and eliminate other streams that belong to the same concept. Then, the retrieval process is done at the receiver using identified implications.

3.5 Threshold-based EEG Transceiver Design

The proposed transceiver enables an adaptive threshold-based compression taking into account the characteristics of the generated traffic, and physical layer components.

3.5.1 EEG Signal Characteristics

We first visualize and analyze the EEG signal in the time and frequency domains in order to understand its properties and obtain the best approach of processing and transmission. A normal continuous EEG signal in the time domain is shown in Figure 3.4. Using frequency domain analysis, we can significantly decrease the amount of data to be transferred. Looking at the generated spectrum shown in Figure 3.5, we observe that it is to some extent sparse, or compressible.[4] Thus, we can efficiently reduce transmission energy consumption for such signal through transmitting only energetic Fourier coefficients, while retrieving original signal at the receiver side.

[4]Here compressible means that the generated spectrum f has a large number of frequencies whose entries (i.e., Fourier coefficients) have magnitudes that are small compared to the norm of f (i.e., the energy of f).

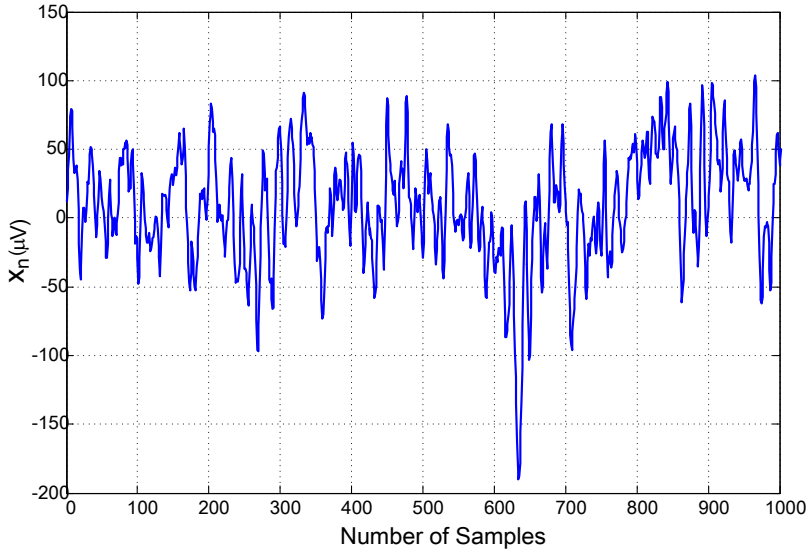


Fig. 3.4 An example of class Healthy EEG signal in the time domain.

3.5.2 Threshold-based Compression

Motivated by the EEG signal characteristics in the frequency domain, we update the OFDM transceiver architecture at the physical layer to support our compression scheme. Unlike the state-of-the-art compression techniques that are applied at the higher layers [44][85], we convey our compression scheme into the physical layer exploiting the existing OFDM transceiver's components in order to perform efficient compression without adding much complexity.

As mentioned, given the basic OFDM transceiver architecture in Figure 3.2, we have added two blocks in order to implement our TBC scheme, namely, the TBC and the FFT Vector Reconstruction. In the TBC block, leveraging the fact that several Fourier coefficients x_f of the EEG signal x have negligible magnitude after IFFT (see Figure 3.6), we consider as 0s all symbols with magnitude lower than a predefined threshold δ (see Figure 3.8). The threshold is set according to the channel characteristics and the maximum distortion that can be tolerated at the receiver side. Clearly, the higher the value of δ, the larger the compression ratio and the resulting distortion. Then, whenever we have a number of consecutive zeros greater than two, the transmitter does not send them, but it notifies the receiver about the length of this sequence and its position in the stream of transferred data. We remark here that efficient techniques like run-length encoding [100] can be leveraged to perform such task.

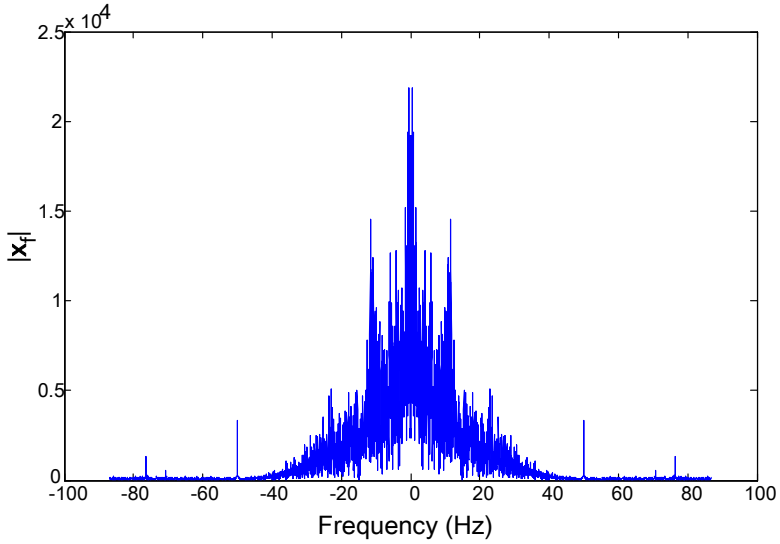


Fig. 3.5 An example of class Healthy EEG signal in the frequency domain.

At the receiver side, the FFT vector reconstruction block is responsible for adding zeros in the received vector at the positions of the ignored symbols before forwarding it to the FFT block. The latter will then demodulate the received symbols and reconstruct the EEG signal.

3.5.3 Error Correction

In order to quantify the achieved compression gain compared to the consequent signal distortion due to our compression scheme, we define the compression ratio as

$$C_r = \left(1 - \frac{\gamma}{\mu}\right) \times 100 \tag{3.9}$$

where γ is the number of data symbols to be transmitted, and μ is the number of the generated data symbols after modulation. While the signal distortion is quantified using PRD, which is given by

$$PRD = \sqrt{\frac{\sum_{i=1}^{N} [x(i) - x_r(i)]^2}{\sum_{i=1}^{N} [x(i) - \bar{x}]^2}} \times 100, \tag{3.10}$$

where $\bar{x}$ is the average value of the original quantized signal, and x_r is the reconstructed one.

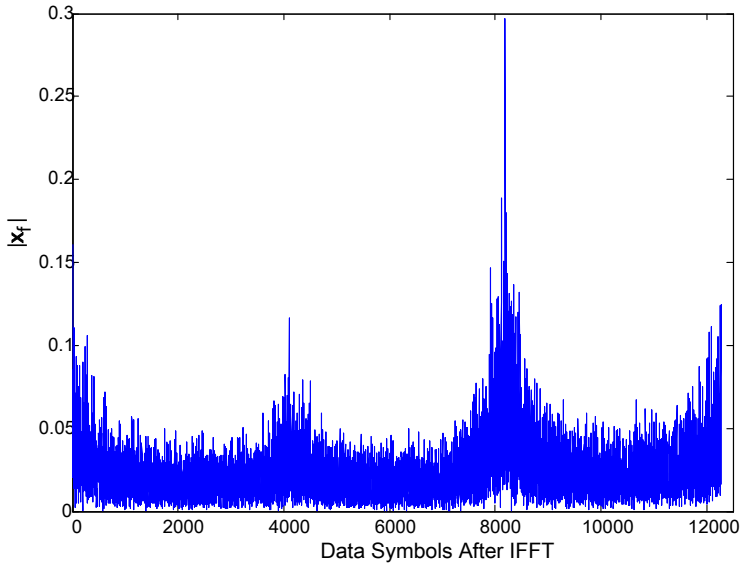


Fig. 3.6 Generated symbols after IFFT while considering 16-QAM modulation.

Interestingly, using our EEG compression transceiver, we can easily define some of the wrong reconstructed samples at the receiver side. As shown in Figure 3.7, some of the wrong samples have very large amplitude compared to the correct samples. This advantage can be used as an Error Correction (EC) scheme in order to decrease Sample Error Rate (SER) and signal distortion at the receiver through: (i) identifying received samples with relatively large amplitude (samples with error), (ii) retransmitting the reconstructed samples with error.

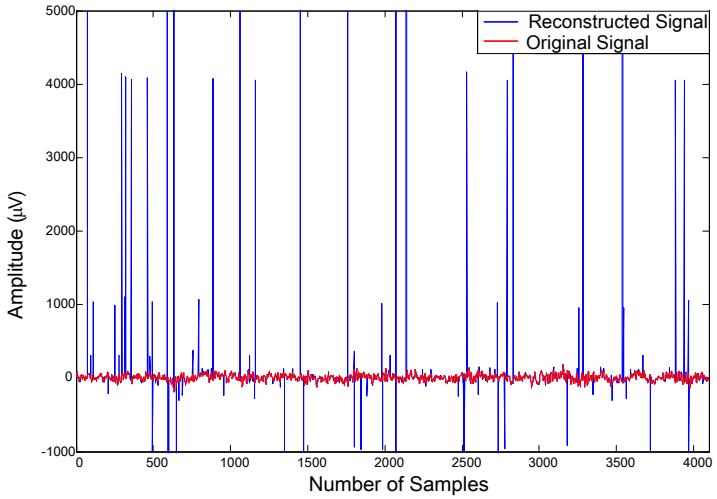


Fig. 3.7 A comparison between reconstructed and original EEG signal for healthy class.

Despite the achieved compression ratio using TBC, we found that it is of prominent importance to further analyze the effect of symbol mapping and modulation on EEG signal characteristics in order to enhance the compression ratio. As noted from Figure 3.5 and Figure 3.6, the EEG signal characteristics after modulation and IFFT modules have been changed and turned to be less compressible. This is mainly due to the effect of symbol mapping and modulation, since representing each data sample with multiple symbols turns the generated symbols after IFFT to be less compressible, i.e., most of the generated symbols after IFFT will have large magnitudes, thus, cannot be neglected. We will tackle this challenge in the next two sub-sections.

3.5.4 Higher-order Modulation

To tackle the problem of symbol mapping effect on EEG characteristics and increase compression efficiency of our transceiver, we study the characteristics of generated symbols after Fourier transform with and without symbol mapping and modulation (see Figure 3.6 and Figure 3.5). We find that exploiting higher-order modulation can help in increasing compression ratio of our transceiver through representing each EEG sample in one symbol, which relieves the effect of symbols mapping. However, as shown in Figure 3.8-(a), magnitudes of the generated symbols after IFFT $|x_f|$ are still less compressible compared to the original case without modulation, i.e., in Figure 3.5 (even after considering the higher-order modulation). As a result, when applying our threshold-based compression, some of the important symbols may be also neglected. To avoid this, we apply *Symbols Masking* before compression. This masking is based on our prior knowledge about the EEG characteristics in the frequency domain. We define a window size W which is the percentage of compressible symbols relative to the total number of symbols. Using this masking, we define the less important symbols of x_f to be passed by the TBC scheme, while isolating more important symbols from compression (see Figure 3.8-(b)). Using such masking with higher-order modulation can significantly mitigate the effect of symbols mapping and modulation on EEG characteristics. By doing so, we could obtain higher compression ratio compared to initial TBC scheme with lower order modulation, as will be shown in simulation results.

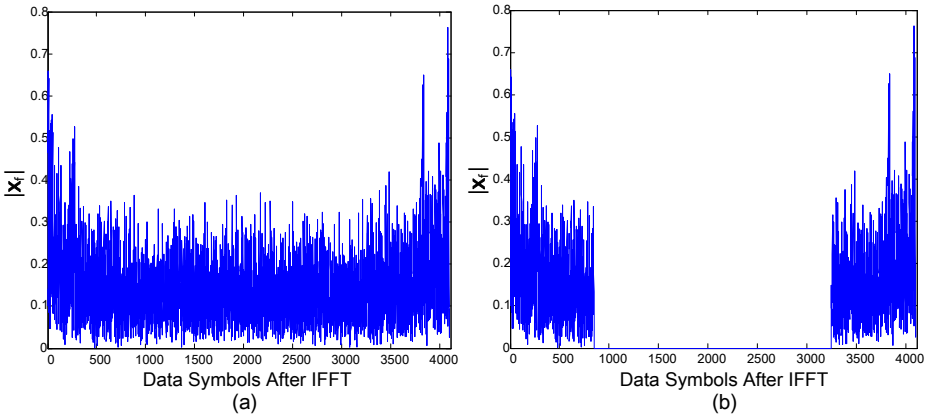


Fig. 3.8 Generated symbols after IFFT while considering 256-QAM modulation, (a) before compression, (b) after compression.

3.5.5 Steam-based Compression

Due to the quality of wireless channel, hardware design, or standards limitations, leveraging higher-order modulation may not be recommended in all cases. Thus, in order to make our transceiver adaptive for different channel conditions and modulation schemes, we propose a Stream-Based Compression (SBC) scheme. Leveraging the generated symbol streams in Section 3.4, the compression ratio can be further increased as follows. The independent streams of symbols are forwarded to the modulation and IFFT blocks, thus at TBC block, we can deal with each stream separately using different values of the threshold δ. This, as also shown in the simulation results section, yields a greater overall compression ratio.

For instance, using QPSK modulation and $L = 12$ bits, we generate 6 streams of symbols. The symbols in each stream will have different values before modulation (see Figure 3.9) and after IFFT (see Figure 3.10). Thus, we can set per-stream thresholds so that each stream can be compressed as much as possible while still meeting the requirement on the maximum allowed distortion.

We remark that discovering the dependency between different streams and selecting only the independent streams is performed before IFFT (i.e., it pertains to the higher-layers of the transceiver architecture), while only the threshold-based compression is done after IFFT, i.e., in the physical layers of the transceiver. Thus, to summarize, the main steps of our SBC scheme are as follows (see Figure 3.11):

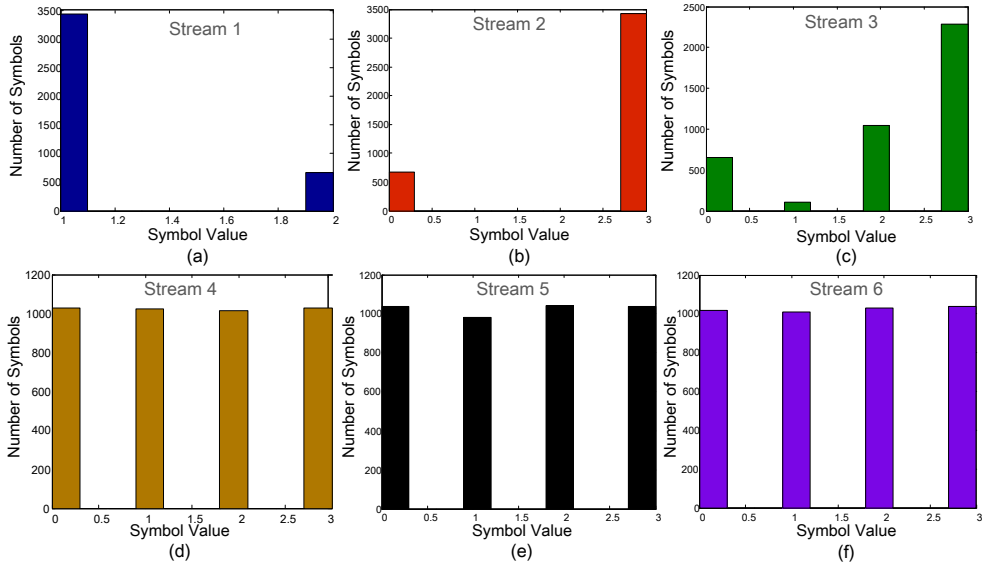


Fig. 3.9 Generated symbols' streams before modulation.

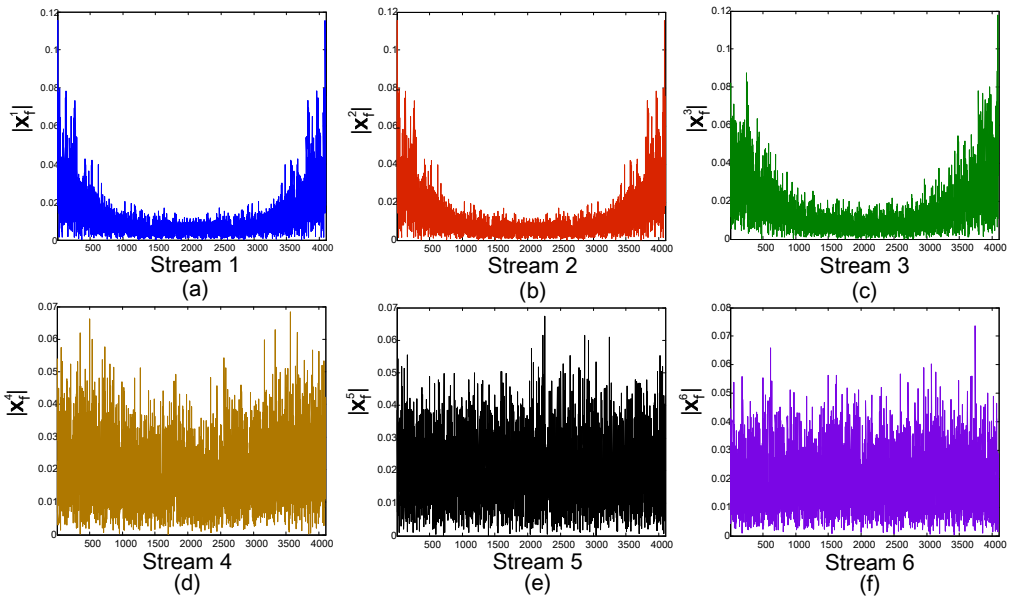


Fig. 3.10 Generated symbols' streams after IFFT while considering QPSK modulation.

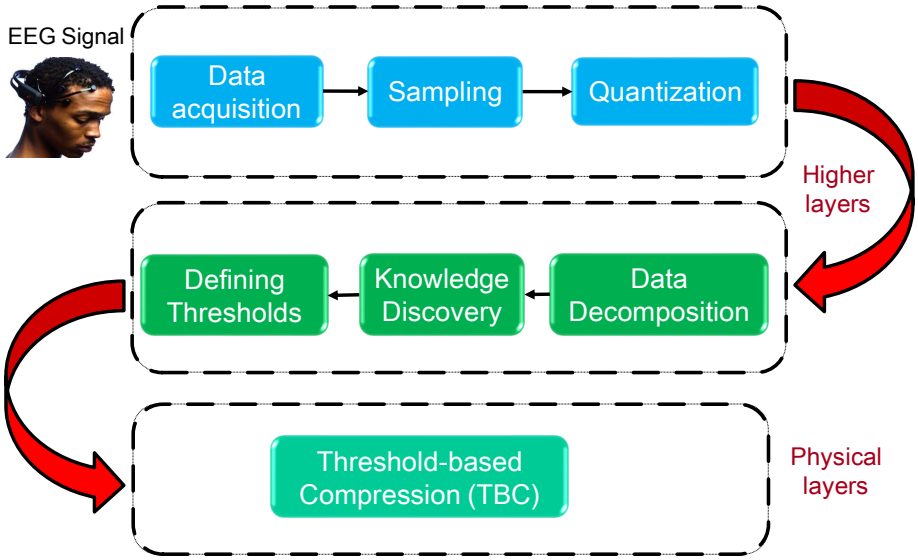


Fig. 3.11 The main steps of the proposed SBC scheme.

- Higher-layers steps, which include stream creation, knowledge discovery, and defining the threshold δ, for individual streams.
- Physical-layer step, which includes TBC.

While at the receiver side, the inverse process is adopted through: (i) using FFT vector reconstruction, which is responsible for adding zeros in the received vector at the positions of the compressed symbols before forwarding it to the FFT, (ii) leveraging obtained dependency between different streams to retrieve discarded streams from transmission.

3.6 Performance Evaluation

In order to derive our simulation results, we consider the system model shown in Figure 3.1 and use the EEG dataset in [51]. Also, to quantify the performance gain provided by our solution, we investigate both the compression ratio and the consequent signal distortion, while considering high signal-to-noise ratio (SNR) for the wireless channel. The simulation parameters we used herein are reported in Table 3.3.

First, we assess the performance of the proposed TBC transceiver in Section 3.5 without performing the signal decomposition into different symbol streams.

Table 3.3 Simulation Parameters

Parameter	T	N_s	T_s	L	M
Value	23.6 sec	4096	0.0058	12 bits	$\in \{2,3,4,6\}$

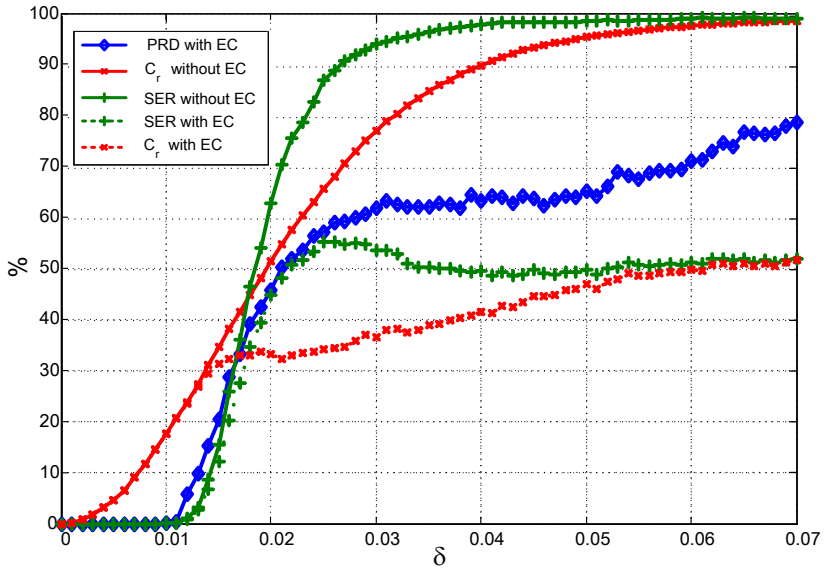


Fig. 3.12 Effect of varying C_r on signal distortion and SER when the TBC scheme and 16-QAM modulation are used.

Figure 3.12 shows the performance gain of the transceiver when the 16-QAM modulation (i.e., $M = 3$ symbols per sample) is used. Herein, we gradually increase the compression ratio C_r by increasing the threshold δ; furthermore, both the cases with and without our Error Correction (EC) scheme are considered. As expected, with increasing δ, the Sample Error Rate (SER) and signal distortion (PRD) increase. However, when EC is applied, SER and PRD reduce significantly thanks to the retransmission of the erroneous samples. On the contrary, the actual or effective C_r decreases due to the higher retransmission overhead. Importantly, these results show that, using the well-known OFDM transceiver architecture with slight modifications, we can obtain about 25% compression ratio while keeping SER and distortion below 10%, which is acceptable by many applications.

Figure 3.13 highlights the increase in C_r that we can obtain by leveraging higher-order modulation and symbols masking. We can now achieve about 40% compression ratio, while keeping distortion around 10%. Also, with larger window size W, the compression ratio grows at the expense of increasing signal distortion. We remark that, depending on the quality of the wireless channel, the modulation order can be

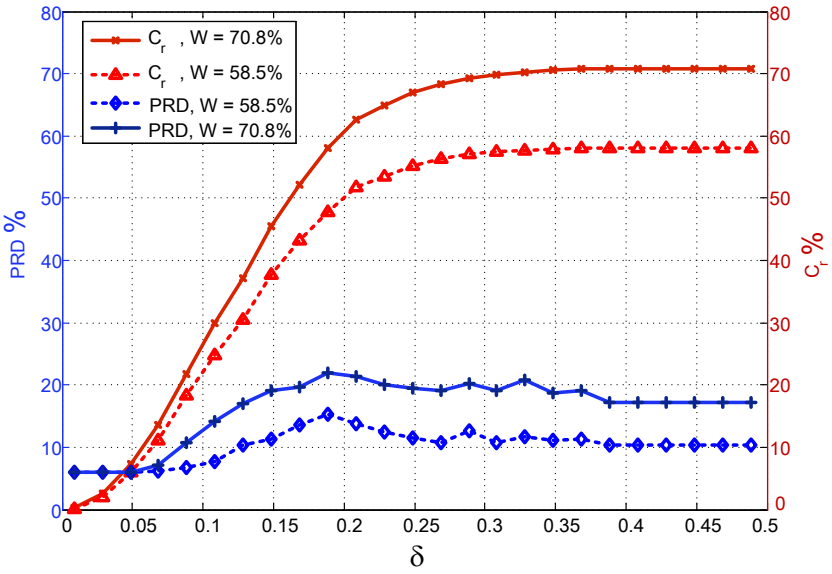


Fig. 3.13 Effect of varying the threshold δ on C_r and signal distortion using TBC scheme and 256-QAM modulation, for different window size W.

increased (i.e., enabling high-order modulations for low channel errors), hence the compression ratio, while still meeting the application requirements.

Next, we assess the performance of the proposed SBC scheme in Section 3.4, i.e., we also account for the benefits brought by the decomposition of the signal into streams of symbols and their processing. Interestingly, our SBC transceiver can support both lossless and lossy compression. As depicted in Figure 3.14, we can achieve about 45% compression ratio at 0% SER and distortion, or about 55% compression ratio at less than 10% SER and distortion. Herein, we used the QPSK modulation with two compression thresholds δ_1 and δ_2, where δ_2 is fixed to 0.011 while δ_1 varies. In particular, δ_2 was used for stream 3, since its values have high variability before modulation and low amplitude after IFFT (see Figure 3.9 and Figure 3.10), while δ_1 was adopted for the other streams. Interestingly, such results show that, thanks to the signal decomposition into streams, we can significantly increase the compression ratio while applying low-order modulation schemes.

The transceiver performance further improves if the Stream-Based Compression with Knowledge Discovery (SBC-KD) scheme is used. Indeed, by applying knowledge discovery and transmitting only the minimal-representation streams, we can considerably reduce the amount of transferred data while still accurately reconstructing the signal at the receiver side. The results in Figure 3.15 demonstrate that in this

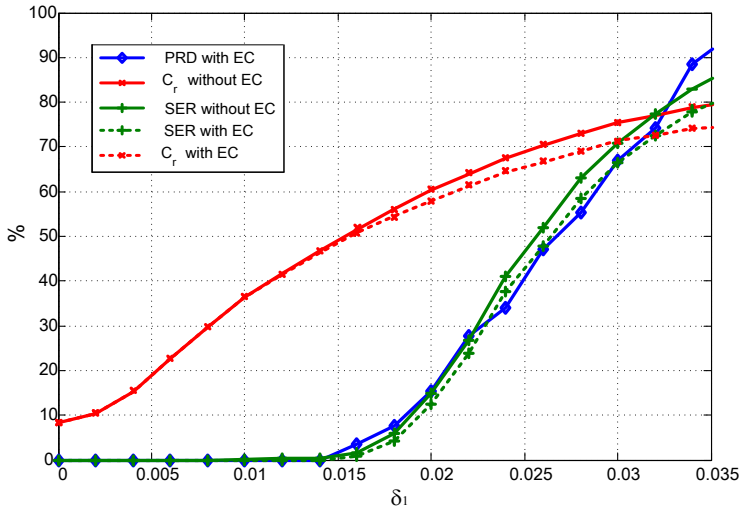


Fig. 3.14 Effect of varying threshold δ_1 on the compression ratio, signal distortion, and SER, when the QPSK modulation is used.

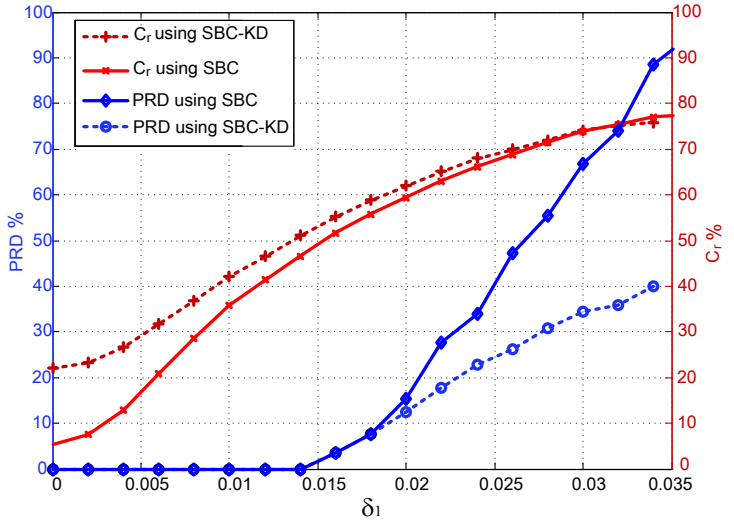


Fig. 3.15 Effect of knowledge discovery on enhancing compression ratio and signal distortion for QPSK modulation.

case we can obtain, roughly, 50% compression ratio at 0% SER and distortion, or 67% compression ratio with less than 20% distortion.

Finally, in Figure 3.16 we compare the performance of the proposed SBC-KD scheme with the DWT technique. Wavelet-based compression techniques consist of transmitting the most significant wavelet coefficients. The strategy adopted to select such coefficients is the main distinguishing factor of the algorithms proposed in the literature [85]. Comparing to DWT-Level thresholding [101], we obtain

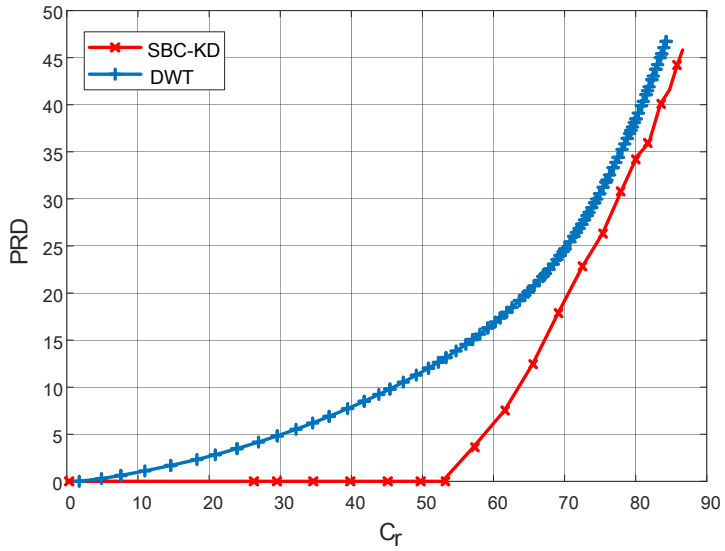


Fig. 3.16 Distortion variation for different values of compression ratio, for the proposed SBC-KD technique and the DWT-level thresholding scheme.

13% reduction in the PRD for compression ratios up to 50%, while achieving 5% reduction in the PRD for higher values, namely, up to 80%, of the compression ratio. Furthermore, we can use the proposed scheme for lossless compression for compression ratios up to 50%, which shows significant gains over DWT in applications requiring zero distortion and high quality analysis of the vital signs.

3.7 Summary

In this chapter, we proposed a novel transceiver design based on symbol-streams compression: the generated symbols are grouped into streams, and only streams that are independent of each other are compressed and transmitted. Additionally, streams are compressed separately, thus the compression thresholds can be tailored to each stream so that the compression ratio is increased while yielding low distortion. In this context, we focused on the case of the EEG signal and showed how the Fourier coefficients representing such signal can be effectively compressed while accounting for the wireless channel characteristics and the application requirements in terms of signal distortion.

Our results show the excellent performance of the proposed design in terms of data reduction gain, signal distortion, low complexity, and the advantages that it

exhibits with respect to state-of-the-art techniques since we could obtain about 50% compression ratio at 0% distortion and sample error rate.

Chapter 4

Distributed In-network Processing and Resource Optimization over Mobile-health Systems

4.1 Overview

This chapter integrates wireless network characteristics and application-layer requirements to provide sustainable, energy-efficient and high-quality services for s-health systems.[1] The proposed scheme enables energy-efficient high-quality patient health monitoring to facilitate remote chronic disease management. In particular, we propose a multi-objective optimization problem that targets different QoS metrics at the application layer like signal distortion, and at the physical layer like transmission delay and Bit Error Rate (BER), as well as monetary cost and transmission energy. In particular, we aim to achieve the optimal tradeoff among the above factors, which exhibit conflicting trends. The main contributions of our work can be summarized as follows:

1. Formulate a cross-layer multi-objective optimization model that aims at adapting and minimizing, at each PDA, the encoding distortion and monetary cost at the application layer, as well as the transmission energy at the physical layer, while meeting the delay and BER constraints.

[1] This work has been published in [102].

2. Leverage geometric program transformation to convert the aforementioned problem into a convex problem, for which an optimal, centralized solution is obtained.

3. Propose a distributed solution through leveraging Lagrangian duality theory. The dual decomposition approach enables us to decouple the problem into a set of sub-problems that can be solved locally, leading to a distributed algorithm that converges to the optimal solution.

4. The proposed distributed algorithm is analyzed and compared to the centralized approach. Our results show the efficiency of our distributed solution, its ability to converge to the optimal solution and to adapt to varying network conditions.

4.2 Related Work

The investigated approaches in the field of m-health can be broadly classified into five categories: energy efficient BASNs design, wireless transmission resource allocation and optimization, implementation of smartphone health monitoring and BCI applications, efficient low-power hardware designs, as well as signal compression, feature extraction, and classification algorithms.

Among different factors, energy efficiency in BASNs, and in general m-health systems, is one of the most challenging problems due to the requirements for high QoS and low transmission delay given the resource constraints. Many of the existing studies focus on Routing, MAC, and physical layer design to address energy and power issues [103]. The basic idea of these techniques is to design new communication methods that obtain optimal performance under the resource constraints. For example, authors in [104] presented a multi-channel MAC protocol (MC-LMAC) that is designed for maximizing system throughput. MC-LMAC combines the advantages of interference-free and contention-free parallel transmissions on different channels. However, the overhead added by this solution is high, and the channel/slot utilization is low for low data rates. The authors in [105] developed a MAC model for BASNs to fulfill the desired reliability and latency of data transmission, while simultaneously maximizing battery lifetime of individual body sensors. In [106], the authors studied the energy-distortion tradeoff from the information-theoretic point of view, in the context of various joint source-channel coding problems.

Wireless transmission optimization in m-health systems has also been widely investigated. For instance, authors in [107] analyzed the relationship between the source rate and the uninterrupted lifetime of a sensor. They formulated a steady-rate optimization problem to minimize rate fluctuation with respect to average sustainable rate. Moreover, they minimized the transmission power of the data aggregators, subject to some power constraints, the requirements on packet loss rate, transmission BER, and packet delay. However, they neither consider the signal processing part in their model nor take the application characteristics into consideration.

In addition to that, the growing power requirements and the need for green communications motivate developing energy efficient techniques to minimize power consumption in next-generation wireless networks, while meeting high user's QoS expectations [108]. In this context, the authors in [109] proposed a hybrid multimedia delivery solution, which achieves an energy-quality-cost tradeoff by combining an adaptive multimedia delivery mechanism with a network selection solution. Based on user preferences, location-based and network related information, the proposed solution in [109] determines whether to adapt multimedia delivery or handover to a new network by computing a score function for each of the selected candidate networks. Then, it selects the network with the highest score as the target network. In [110], the authors focused on the energy efficient design of physical-layer transmission technologies and MAC-layer radio resource management. They studied the tradeoff between spectrum efficiency and energy efficiency as part of their optimization model. Some studies have also focused on joint compression and communication optimization, where the compression power consumption and transmission power consumption are jointly considered in order to optimize the performance of the entire system. However, this approach is mainly applied to video transmission systems, since the video encoding itself consumes high power compared with the wireless transmission [111]. In general, it is agreed that energy-efficient cross-layer design is a very complex problem, since it requires to effectively investigate all the network layers optimizations jointly [112].

The immense advancements in smartphone features and capabilities have promoted the development of smartphone apps for long-term chronic condition management. Health-related smartphone apps can build a sense of security for patients with chronic conditions, since they felt secure that their states are carefully monitored, and their doctors take care of them even outside the hospital or clinic. Thus, there is a growing interest in the literature in leveraging mobile apps to enhance healthcare

services for chronically ill and elderly people. For instance, the authors in [113] have implemented an embedded low cost, low power web server for internet based wireless control of BCI-based home environments. This web server provides remote access to the environmental control module through transmitting BCI output commands determined by BCI system to drive the output devices. In [114], the authors presented a real-time mobile adaptive tracking system, where the wireless local area network, or third-generation-based wireless networks are used to transfer test results from a smartphone to the remote database. This system provides real-time classification of test results, generation of appropriate short message service-based notification, and sending of all data to the Web server. We remark here that aspects related to wireless transmission, channel characterization, and transmission energy minimization are not within the scope of this previous work. A comprehensive overview of recent smartphone apps designed for remote health monitoring can be found in [115][116]. However, more studies involving larger samples size, patients, and health professionals are still necessary to investigate mobile apps' acceptability and effectiveness.

Accordingly, studying energy and monetary cost minimization in wireless transmission, as well as signal distortion tradeoff for delay sensitive transmission of medical data should be taken into consideration. However, neither the aforementioned work nor the studies in [117] and [118], have considered a cross-layer approach that takes the application requirements, in-network data processing, and physical layer components jointly into consideration. With regards to our previous work [119][120][121], we have studied the transmission and processing energy consumption and developed an Energy-Compression-Distortion analysis framework. Using this framework, [119] proposed a cross-layer optimization model that minimizes the total energy consumption, under a TDMA scheduling. The work has been extended in [120] to the case where more than one link can be activated simultaneously, using the same TDMA slot. Furthermore, to evaluate and verify our model, we have developed a smartphone app in [121], where the PDA compresses the gathered EEG data using dynamically obtained optimal compression parameters based on real-time measurements of the packet delivery ratio and end-to-end delay, then forwards it to the healthcare server which decompresses and reconstructs the original signal. However, this previous work addresses energy consumption minimization only using centralized approach, and it completely ignores the energy-cost-distortion tradeoff.

Finally, [122] presents a preliminary version of our study, where only one single PDA is considered.

4.3 System model and objectives

First, we start by introducing the system under study, as well as the application and network requirements, which will be addressed in our optimization problem.

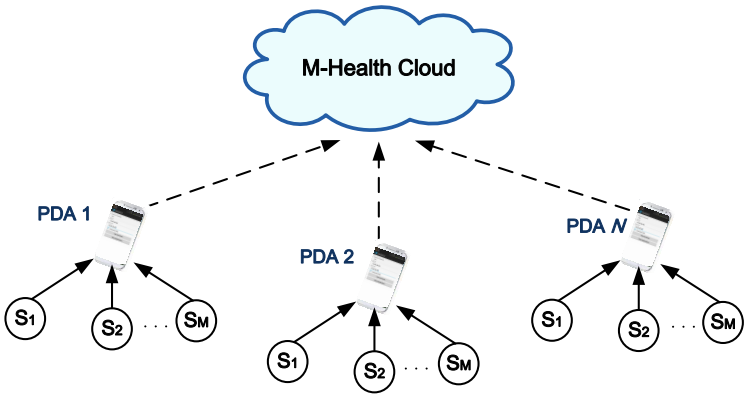


Fig. 4.1 System Model.

4.3.1 Reference scenario

We consider a wireless m-health system, as shown in Figure 4.1. In our model, PDA gathers sensed data from a group of sensor nodes, compresses it, and then forwards the aggregate traffic to the MHC. The MHC can be considered as a virtual central node that is responsible for gathering the transmitted data from PDAs, and for coordinating different PDAs in a central fashion whenever needed. In addition to that, signal reconstruction, feature extraction, classification and distortion evaluation can be performed at the MHC to detect the status of the patient [123]. We denote the number of PDAs that want to transfer their data to the MHC by $\tilde{N}$. Each PDA will perform DWT compression, quantization and encoding on the raw EEG data, and it will transmit the output through its RF interface [124] (see Figure 4.2). It is assumed that the PDA can adapt the transmission power level as well as the compression ratio, according to, e.g., the radio propagation conditions. In particular, the PDA employs a threshold-based DWT so that the number of output samples generated from DWT,

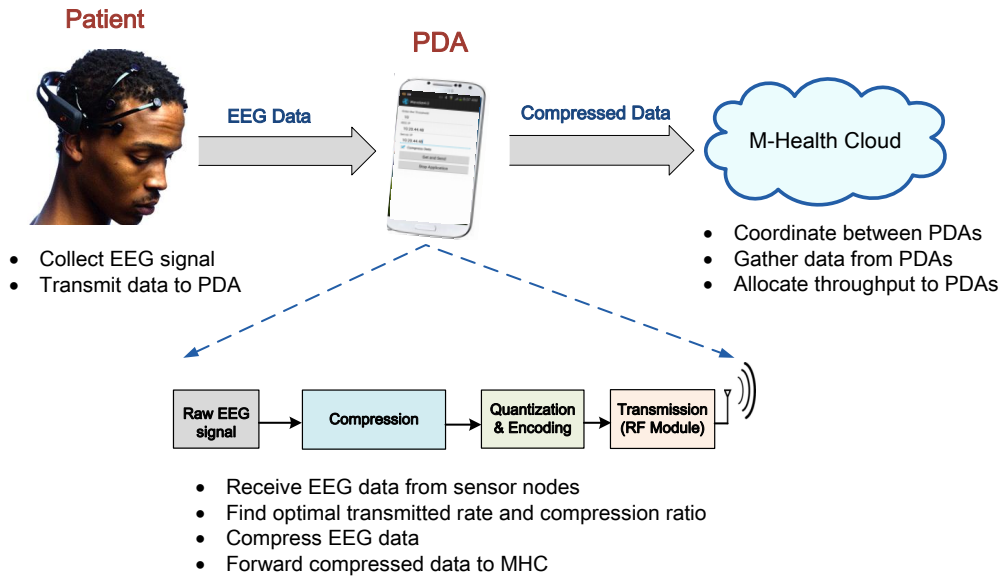


Fig. 4.2 A detailed medical EEG system diagram.

and thus the compression ratio of the DWT, can be easily controlled. Although the adopted framework employs the encoding model of EEG signal, it can be easily extended to different data types which are typically at a low data rate (e.g., body temperature, blood pressure or heart-rate reading), or at higher data rates such as video streaming.

Furthermore, each PDA receives from the sensor nodes (i.e., S_1 to S_M), the application layer constraints, namely, maximum BER and data transfer delay. After that, given the channel conditions, it determines the transmitted rate and compression ratio that satisfy the application layer constraints, while providing the optimal tradeoff among energy consumption, monetary cost, and signal distortion.

4.3.2 Performance metrics

Following [125], we express the QoS requirements of our healthcare application through signal distortion, BER, and data transfer delay from the PDA to the MHC. In particular, typically BER and transfer delay are constrained not to exceed a given maximum value (i.e., $\vartheta = 10^{-6}$ and $\tau = 10$ ms, respectively), while the distortion D_i with which the signal from the generic PDA i ($i = 1,\ldots,\tilde{N}$) is reconstructed at the receiver side should be as small as possible. Using the results obtained through

our real-time implementation [121], such quantity can be written as:

$$D_i = \frac{c_1 e^{(1-\kappa_i)} + c_2 \cdot (1-\kappa_i)^{-c_3} + c_4 \cdot F^{-c_5} - c_6}{100}. \tag{4.1}$$

where F is the wavelet filter length [68], κ_i is the compression ratio, and the model parameters c_1, c_2, c_3, c_4, c_5 and c_6 are estimated by the statistics of the typical EEG encoder used in [121]. We remark that given a string of l_s bits representing raw EEG samples, the encoder at PDA i will output $l = l_s(1-\kappa_i)$ bits to be transmitted through the radio interface.

Then, looking at the communication network, it is of paramount importance to minimize the energy consumption and the monetary cost that the PDA incurs for transferring its data to the MHC. We define the energy expenditure of PDA i to transmit l_i bits over a channel with bandwidth w, at rate r_i, as [120]:

$$\widetilde{E}_i = \frac{l_i \cdot N_0 \cdot w}{r_i \cdot g_i}(2^{r_i/w} - 1) \tag{4.2}$$

where N_0 is the noise spectral density and the channel gain g_i is given by

$$g_i = k \cdot \alpha \cdot |h_i|^2. \tag{4.3}$$

In (4.3), $k = -1.5/[\log(5\vartheta)]$, α is the path loss, and $|h_i|$ is the fading channel magnitude. Also, we remark that the above equations express the relationship between energy consumption and BER: the lower the BER, the higher the energy that is required for data transmission.

The monetary cost to send l_i bits is instead expressed in Euro and defined as:

$$\widetilde{C}_i = \varepsilon \cdot l_i \tag{4.4}$$

where ε is the cost of sending one bit. Monetary cost could be easily obtained through the use of IEEE 802.21 standard [126], which allows a user device to gather information about the available wireless networks (of course, such value can be updated if there are any changes in pricing) [127].

Importantly, looking at the above expressions, it can be seen that some factors exhibit opposite trends. The higher the compression ratio κ_i, hence, the signal distortion, the fewer the transmitted bits (l_i). Conversely, the smaller the l_i, the lower

the energy consumption and the monetary cost. Thus, in order to achieve the system goals stated above, it is necessary to find the optimal tradeoff between two conflicting objectives, energy and cost on one side and distortion on the other. We take this challenge in the next section, where our optimization problem is set out.

4.4 Energy-Cost-Distortion Optimization

As mentioned, the proposed optimization problem considers three criteria: transmission energy consumption, monetary cost, and encoding distortion. Each criterion presents different ranges and units of measurement, thus we need first to normalize them in order to make them adimensional and comparable. Considering that distortion is already expressed as a percentage, we normalize $\widetilde{E}_i$ and $\widetilde{C}_i$ with respect to their maximum value, i.e., the value they take when no compression is used (i.e., $\kappa_i = 0$). We denote the normalized energy and monetary cost (hereinafter referred to as cost for brevity) by E_i and C_i.

Then, considering that the problem can be solved in a centralized manner at the MHC, we write our objective function as:

$$\min_{\kappa_i, r_i} \sum_{i=1}^{\tilde{N}} \lambda \cdot (E_i + C_i) + (1-\lambda) \cdot D_i \tag{4.5}$$

where λ is a weighting factor, $0 < \lambda < 1$, that can be set by the PDA based on the desired energy-cost-distortion tradeoff. In particular, $\lambda = 1$ means that we aim at minimizing the transmission energy and monetary cost only, and neglect the distortion. On the contrary, $\lambda = 0$ means to neglect transmission energy and monetary cost, and only consider distortion. The unknowns in this problem are the transmission rates r_i, on which E_i depends, and the compression ratios κ_i, on which all three performance metrics (E_i, C_i and D_i) depend. Indeed, recall that, given the number l_s of raw bits as input to the encoder of PDA i, the number of output bits to be transmitted is given by $l = l_s(1-\kappa_i)$.

Based on the requirements of the healthcare applications and of the communication networks, the above expression should be minimized subject to the following

constraints:

$$\frac{l_s(1-\kappa_i)}{r_i} \leq \tau, \qquad \forall i \in \tilde{N} \tag{4.6}$$

$$\sum_{i=1}^{\tilde{N}} r_i \leq B \tag{4.7}$$

$$0 \leq r_i, \;\; 0 \leq \kappa_i \leq 1, \qquad \forall i \in \tilde{N} \tag{4.8}$$

Constraint (4.6) accounts for the fact that the transmitted data must be received at the MHC with a maximum transmission delay τ, which for simplicity we assume to be the same for all PDAs. The network perspective instead is accounted for by constraint (4.7), which states that sum of the transmission rates of the PDAs cannot exceed the maximum available bandwidth B. Finally, constraint (4.8) simply ensures that the decision variables take non-negative values.

Unfortunately, this initial form of the optimization problem is non-convex [128]. One of the common methods to make the problem convex, is to transform the original problem into a Geometric Programming (GP) problem [129]. For the transmission energy in (4.2), we can use the Taylor Series Expansion, $2^{r_i/w} = 1 + \frac{r_i \log(2)}{w} + \frac{{r_i}^2 \log^2(2)}{2w^2} + \frac{{r_i}^3 \log^3(2)}{6w^3} + \frac{{r_i}^4 \log^4(2)}{24w^4} + O({r_i}^5)$. Then, the objective function can be transformed into an equivalent convex one using a change of variables. Define $\widehat{\kappa_i} = \log(1-\kappa_i)$, and $\widehat{r_i} = \log(r_i)$. By substituting these expressions in (4.2), we have:

$$\widehat{E_i} = \frac{l_s(1-\kappa_i)w\left(\frac{r_i \log(2)}{w} + \frac{{r_i}^2 \log^2(2)}{2w^2} + \cdots\right)N_0}{r_i \cdot g_i} \tag{4.9}$$

$$= \frac{l_s(1-\kappa_i)\left(\log(2) + \frac{r_i \log^2(2)}{2w} + \cdots\right)N_0}{g_i} \tag{4.10}$$

$$= \frac{l_s e^{\widehat{\kappa_i}}\left(\log(2) + \frac{e^{\widehat{r_i}} \log^2(2)}{2w} + \cdots\right)N_0}{g_i}. \tag{4.11}$$

Using the same approach, we will have

$$\widehat{D_i} = c_1 e^{e^{\widehat{\kappa_i}}} + c_2 e^{-c_3 \widehat{\kappa_i}} + c_4 \cdot F^{-c_5} - c_6 \tag{4.12}$$

, and

$$\widehat{C}_i = \varepsilon \cdot l_s e^{\widehat{\kappa}_i}. \tag{4.13}$$

Then, we write the optimization problem as:

$$\begin{aligned}
&\min_{\widehat{\kappa}_i, \widehat{r}_i} \log \sum_{i=1}^{\tilde{N}} U_i(\widehat{\kappa}_i, \widehat{r}_i, \lambda) \\
&\text{such that} \\
&\log(l_s \cdot e^{\widehat{\kappa}_i - \widehat{r}_i}) \leq \log \tau, \qquad \forall i \in \tilde{N} \\
&\log\left(\sum_{i=1}^{\tilde{N}} e^{\widehat{r}_i}\right) \leq \log B \\
&\widehat{\kappa}_i \leq 0, \qquad \forall i \in \tilde{N}
\end{aligned} \tag{4.14}$$

where

$$U_i(\widehat{\kappa}_i, \widehat{r}_i, \lambda) = \lambda \cdot (\widehat{E}_i + \widehat{C}_i) + (1-\lambda) \cdot \widehat{D}_i \tag{4.15}$$

is a sum of exponential functions, and the objective function in (4.14) is now convex in $\widehat{\kappa}_i$ and $\widehat{r}_i$. Thus, using the centralized approach, several efficient solution methods can be applied to solve this problem. The globally optimal solution to the original optimization problem can be obtained by solving (4.14) and then computing $\kappa_i{}^* = 1 - \exp(\widehat{\kappa}_i^*)$, and $r_i^* = \exp(\widehat{r}_i^*)$, with $\widehat{\kappa}_i^*$ and $\widehat{r}_i^*$ being the optimal solution for (4.14). At last, we note that in the above problem, the number of variables grows as $2\tilde{N}+1$, while the number of constraints grows as $3\tilde{N}+2$.

4.5 Distributed solution

In large scale networks and heterogeneous m-health systems, the above centralized optimization becomes inefficient and quite complex. Indeed, solving the minimization problem that we have formulated in a centralized fashion requires that global information about the network is available at the MHC, and, in many cases, the overhead due to such communication from the PDAs to the MHC cannot be sustained. Thus, in the following we formulate the dual problem of (4.14) and decompose the original problem into smaller sub-problems, which can be efficiently solved in a distributed fashion while still achieving optimality. Note that we also provide an iterative algorithm that allows the MHC to optimally set λ when the best tradeoff

between energy and monetary cost on one hand and distortion on the other should be established.

4.5.1 Dual decomposition

Convex optimization has highly-useful Lagrange duality properties, which leads to decomposable structures. Lagrange duality theory adapts the original minimization problem in (4.14), the so-called *primal problem*, into a *dual problem*. The basic idea in Lagrange duality is to relax the original problem by moving the constraints into the objective function in the form of a weighted sum. The Lagrangian of (4.14) is defined as

$$L(\widehat{\kappa},\widehat{r},\mu,\nu) = \log\left(\sum_{i=1}^{\tilde{N}} U_i(\widehat{\kappa}_i,\widehat{r}_i,\lambda)\right) + \sum_{i=1}^{\tilde{N}} \mu_i f_i(\widehat{\kappa}_i,\widehat{r}_i) + \nu h(\widehat{r}_i) \tag{4.16}$$

where $\widehat{\kappa} = [\widehat{\kappa}_1,\ldots,\widehat{\kappa}_{\tilde{N}}]$, $\widehat{r} = [\widehat{r}_1,\ldots,\widehat{r}_{\tilde{N}}]$ and $\mu = [\mu_1,\ldots,\mu_{\tilde{N}}]$. μ_i and ν are the *Lagrange multipliers* related to the i-th inequality constraint $f_i(\widehat{\kappa}_i,\widehat{r}_i) \geq 0$ and the network constraint $h(\widehat{r}_i) \geq 0$, respectively, with

$$f_i(\widehat{\kappa}_i,\widehat{r}_i) = \log\tau - \log(l_s \cdot e^{\widehat{\kappa}_i - \widehat{r}_i}) \tag{4.17}$$

$$h(\widehat{r}_i) = \log B - \log\left(\sum_{i=1}^{\tilde{N}} e^{\widehat{r}_i}\right). \tag{4.18}$$

The dual objective $g(\mu,\nu)$ is defined as:

$$g(\mu,\nu) = \inf_{\widehat{\kappa},\widehat{r}} L(\widehat{\kappa},\widehat{r},\mu,\nu). \tag{4.19}$$

When the problem is convex, the difference between the optimal primal objective U^* and the optimal dual objective g^* reduces to zero [128],[130]. Hence, the primal problem (4.14) can be equivalently solved by solving the dual problem

$$\max_{\mu,\nu} \quad g(\mu,\nu) \tag{4.20}$$

$$\text{s.t.} \quad \nu \geq 0,\ \mu_i > 0 \forall i. \tag{4.21}$$

Since $g(\mu, \nu)$ is differentiable, the master dual problem can be solved with the gradient method [131], where the dual variable ν at the $(t+1)$-th iteration is updated by

$$\nu(t+1) = \nu(t) + \beta \frac{\partial L}{\partial \nu} \tag{4.22}$$

with $\beta > 0$ being the gradient step size. The gradient method is guaranteed to converge to the optimal value as long as the step size is sufficiently small [130]. Given the dual variables at the t-th iteration, the primal variables $\widehat{\kappa}_i^*$ and $\widehat{r}_i^*$ can be computed by solving the following equations, involving the gradient of L with respect to the Lagrange multipliers and the primal variables:

$$\frac{\partial L}{\partial \widehat{\kappa}_i} = 0 \quad ; \quad \frac{\partial L}{\partial \widehat{r}_i} = 0 \quad ; \quad \frac{\partial L}{\partial \mu_i} = f_i(\widehat{\kappa}_i, \widehat{r}_i) = 0\,. \tag{4.23}$$

4.5.2 Distributed algorithm with fixed λ

Let us assume that the value of λ is predefined by the users according to their preferences. Looking at (4.16) and (4.23), we can see that the Lagrangian can be divided into $\tilde{N}$ separate sub-problems, one for each PDA in the network. The sub-problems can be locally and independently solved provided that ν, i.e., the Lagrange multiplier that is related to the maximum available bandwidth in the network, is known. Thus, we devise an iterative, distributed algorithm, named DOA, that lets each PDA solve its corresponding sub-problem and send to the MHC its optimal values for $\widehat{\kappa}_i$ and $\widehat{r}_i$, while the MHC updates the dual variable ν according to (4.22). The pseudocode of DOA is reported in Algorithm 2.

Initially, each PDA i assumes $\nu = 0$ and computes the dual variable μ_i and the primal variables, i.e., (i) $\widehat{\kappa}_i$, hence the compression ratio κ_i, and (ii) $\widehat{r}_i$, hence the transmission rate r_i. If the bandwidth constraint is satisfied (i.e., $h(\widehat{r}_i) \geq 0 \; \forall i$), the MHC instructs to the PDAs to transmit their data using the calculated compression ratio and transmission rate. Otherwise, the MHC updates the value of ν using (4.22). As ν increases, each PDA will have to decrease its r_i so as to meet the available bandwidth constraint. On the contrary, it will have to increase its κ_i, hence the signal distortion (see Figure 4.3). When the dual variables converge, the primal variables also converge to their optimal values in slightly more than 15 iterations, as shown in Figure 4.4.

Algorithm 2 Distributed Optimization Algorithm (DOA)

1: $t = 0$, $\nu(t) = 0$.
2: Each PDA locally solves its problem by computing the equations in (4.23), and then sends the solution $\widehat{\kappa}_i$, $\widehat{r}_i$, and μ to the MHC.
3: **while** all κ_i's $\leq 1 \wedge h(\widehat{r}_i) < 0 \wedge t < n_{iter}$ **do**
4: The MHC updates ν as in (4.22) and broadcasts the new value $\nu(t+1)$
5: The MHC gets the new estimated parameters $\widehat{\kappa}_i, \widehat{r}_i$ and μ_i from each PDA and computes $h(\widehat{r}_i)$
6: t++
7: **end while**
8: **if** all κ_i's $\leq 1 \wedge h(\widehat{r}_i) \geq 0$ **then**
9: The MHC instructs the PDAs to use the new values for $\widehat{\kappa}_i$ and $\widehat{r}_i$.
10: **else**
11: Break % No feasible solution reached
12: **end if**

4.5.3 Distributed algorithm with varying λ

Now, we focus on the impact of λ on the tradeoff between energy, cost, and signal distortion, that is achieved when the aforementioned problem is optimally solved. As expected, Figure 4.5(a) shows that high values of λ lead to a greater reduction of transmission energy and cost as they are assigned a higher weight, while low λ's provide little distortion at the expense of the transmission energy and cost. Here, we aim at studying the case of great practical relevance where energy consumption, cost, and distortion are all equally important. In this case, it has been shown [132, 133] that the best tradeoff can be obtained by selecting the value of λ maximizing the minimum value of the objective function (i.e., $\lambda = 0.6$ in Figure 4.5(b)).

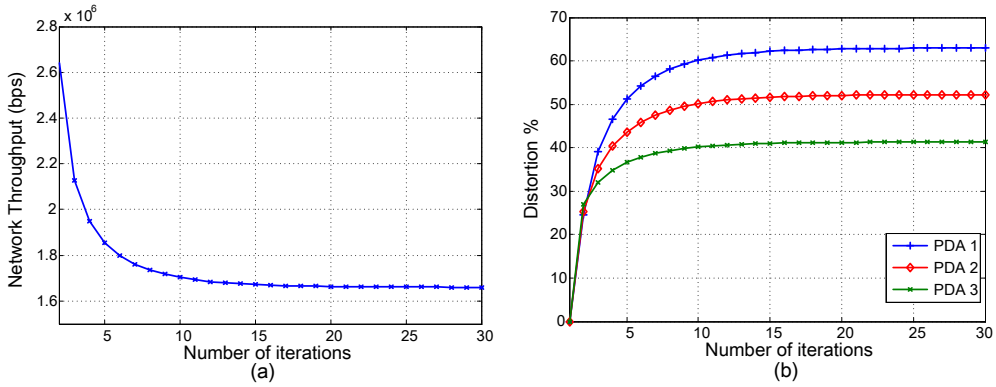


Fig. 4.3 (a) Network throughput and (b) distortion as a function of the number of iterations.

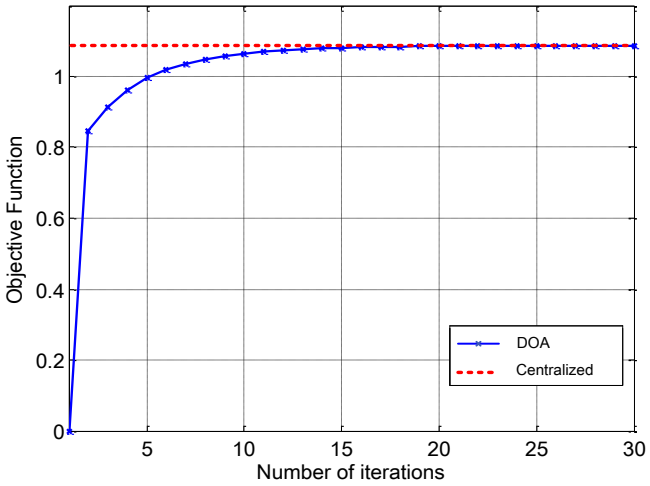


Fig. 4.4 The value of the primary objective function obtained through the DOA algorithm, as the number of iterations increases. The results obtained through the DOA distributed algorithm are compared against the optimal value derived through the centralized solution of the primal problem.

It is easy to see that, if we let the users autonomously determine the value of λ, in some cases the DOA algorithm cannot converge to the optimal solution. Furthermore, even if all users agreed on the same value of λ, we would not have any guarantee that a feasible problem solution exists. As an example, if the users care mostly about the quality level of their reconstructed signal, then distortion will be weighted very high by all of them, likely making the total requested bandwidth exceed the available bandwidth B. We therefore let the MHC determine the value of the weighting factor λ according to the following algorithm, called Algorithm λ-DOA.

We start by looking at the original problem in (4.5) and note that $\lambda = 0$ corresponds to accounting for distortion only. In this case the value of compression factor will be set to the minimum possible, given the maximum allowed data rates and the delay constraint. As λ increases, energy and cost will be weighted more, thus leading to an optimal solution of the problem that requires lower data rates. Based on these observations, we can avoid using ν in the dual problem solution and exploit λ instead. In particular, we can first search for the minimum λ for which the problem is feasible, i.e., the bandwidth constraint is met. Then, we can increase λ so as to find the optimal tradeoff between the term accounting for energy consumption and monetary cost, and distortion, i.e., the optimal λ^* that meets the max-min principle

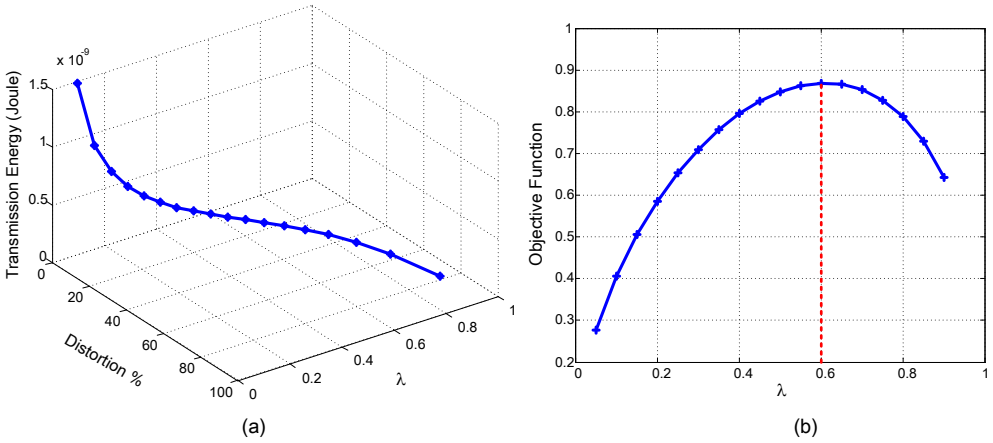


Fig. 4.5 (a) Tradeoff between transmission energy and distortion and (b) value of the minimized objective function, as λ varies.

[132] is found.

$$\lambda^* = \arg\max_{\lambda} \left(\min_{\widehat{\kappa}_i, \widehat{r}_i} \log \sum_{i=1}^{\tilde{N}} U_i(\widehat{\kappa}_i, \widehat{r}_i, \lambda) \right). \tag{4.24}$$

It is important to remark that this approach leads to the same solution as the one obtained by using v. In addition, it allows us to limit the range of possible values of λ to those that make the problem feasible, thus greatly reducing the number of required iterations.

Algorithm 3 details the first step of the procedure. All PDAs start with $\lambda = 0$ and solve the sub-problems locally, thus deriving $\widehat{\kappa}_i$ and $\widehat{r}_i$. If $h(\widehat{r}_i) < 0$ and the maximum number of iterations has not been exceeded, the MHC increases the value of λ, i.e., it assigns more weight to the transmission energy and the monetary cost at the expense of distortion. As a result, eventually the bandwidth constraint (i.e., $h(\widehat{r}_i) \geq 0$) will be satisfied. In order to reduce the number of iterations, the MHC can update the value of λ as follows:

$$\lambda(t+1) = \lambda(t) + \beta \left(\sum_{i=1}^{N} e^{\widehat{r}_i} - B \right) \tag{4.25}$$

where $\beta > 0$ being the gradient step size [131]. We stress that, by doing so, the value of λ computed by the MHC in the first step is the minimum value that satisfies the $h(\widehat{r}_i)$ constraint. After that, the MHC runs Algorithm 4 to find the optimal value λ^*, which maximizes the minimum value of function U.

Algorithm 3 λ-DOA - First step

$\lambda = 0$. **At MHC:**
Get estimated parameters $\widehat{\kappa}_i, \widehat{r}_i$ from each PDA and compute $h(\widehat{r}_i)$
j=0
while $\lambda(j) \leq 1 \wedge h(\widehat{r}_i) < 0 \wedge j < n_{iter}$ **do**
    Compute $\lambda(j)$ using (4.25) and broadcast it to PDAs
    Get new estimated parameters $\widehat{\kappa}_i, \widehat{r}_i$ from each PDA and compute $h(\widehat{r}_i)$
    j++
end while
if $h(\widehat{r}_i) \geq 0$ **then**
    Broadcast λ that ensures the bandwidth constraint is met
    Run Algorithm 4
else
    Break % No feasible solution has been reached
end if

At PDAs:
Receive $\lambda(j)$ from MHC
Solve the equations in (4.23)
Send estimated $\widehat{\kappa}_i, \widehat{r}_i$ to the MHC
if λ^* is received **then**
    Use the estimated $\widehat{\kappa}_i$ and $\widehat{r}_i$ to transmit medical data
end if

Both algorithms require that at each iteration some values are exchanged between PDAs and MHC: specifically, the MHC broadcasts the value of λ to the PDAs while the PDAs send back their estimated optimal values for $\widehat{\kappa}_i$ and $\widehat{r}_i$. The MHC checks the value of $h(\widehat{r}_i)$, and computes λ by solving (4.25), and then feeds them back to the PDAs. Once convergence is reached, each PDA transmits its traffic stream of medical data to the MHC at the optimal transmission rate r_i with the optimal compression ratio κ_i (see Figure 4.6).

Note that our operating environment changes over time: some PDAs may join or leave the network, or the radio propagation conditions may vary, thus the energy consumption per transferred bit as well as the network achievable throughput may change as well. As shown in the next section, in this case, our distributed algorithm are able to readily adapt to the network dynamics, by letting the PDAs and the MHC quickly update the system parameters so as to reach the optimal operational point.

Algorithm 4 λ-DOA - Second step

t=0
Compute $U(t) = \sum_{i=1}^{\tilde{N}} U_i(\widehat{\kappa}_i, \widehat{r}_i, \lambda)$
while $\lambda(t) < 1$ **do**
  $\lambda(t+1) = \lambda(t) + \beta$
  Get new estimated $\widehat{\kappa}_i, \widehat{r}_i$ from each PDA
  if $U(t) < U(t-1)$ **then**
    $\lambda^* = \lambda(t-1)$
    $U^* = U(t-1)$
    Break
  else
    t++
  end if
end while
Broadcast the results to PDAs

4.6 Performance Evaluation

In this section, we first present the network scenario that we used to derive our numerical results. Then, we show the system performance in the case where only one PDA has to transfer data toward the MHC, as well as when multiple PDAs are involved.

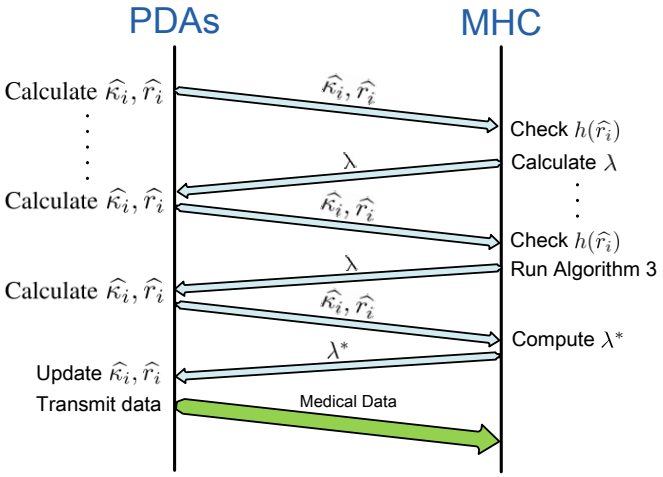


Fig. 4.6 The λ-DOA sequence diagram.

4.6.1 Simulation setup

The generated results were obtained using the system model shown in Figure 4.1, while considering the technical requirements of the selected BASN applications [134]. In the case of multiple PDAs, as an example, we consider 3 PDAs, however, the proposed scheme can be adapted easily to any change in the network topology by adding or removing new PDAs, as will be shown later. Each PDA can capture 173.6 samples of the EEG signal per second, and we assume that samples are collected for 23.6 seconds, corresponding to 4096 samples of epileptic EEG data [51]. Each raw sample is represented using 12 bits. The available bandwidth B is set to 4 Mbps. At the server side, the EEG feature extraction, classification and distortion evaluation can be performed. The target BER is set to $\vartheta = 10^{-6}$. Moreover, to model small scale channel variations, flat Rayleigh fading is assumed, with Doppler frequency of 0.1 Hz. Other simulation parameters are reported in Table 4.1.

Table 4.1 Simulation Parameters

Parameter	**Value**	**Parameter**	**Value**
N_0	-174 dBm/Hz	τ	10 ms
ε	10^{-6} Euro/bit	w	0.5 MHz
F	2	l_s	62 KB
c_1	1.48	c_2	4.35
c_3	1.46	c_4	2.4
c_5	0.18	c_6	9.5

4.6.2 Single-PDA scenario

Here, we investigate the performance of our scheme in the presence of a single PDA that has to send its data to the MHC, while achieving an optimal energy-cost-distortion tradeoff. Figure 4.7 shows the energy-distortion tradeoff, and the monetary cost-distortion tradeoff, for $\lambda = 0.5$. The behavior depicted in the plot, although expected, underscores how critical it is to optimally set the system parameters so as to achieve a good operational point. Indeed, the increase of the compression ratio leads to decreasing the number of transmitted bits, which results in decreasing transmission energy and monetary cost at the expense of increasing signal distortion. Thus, a decrease in the energy consumption and in the monetary cost may have severe effects on the signal distortion.

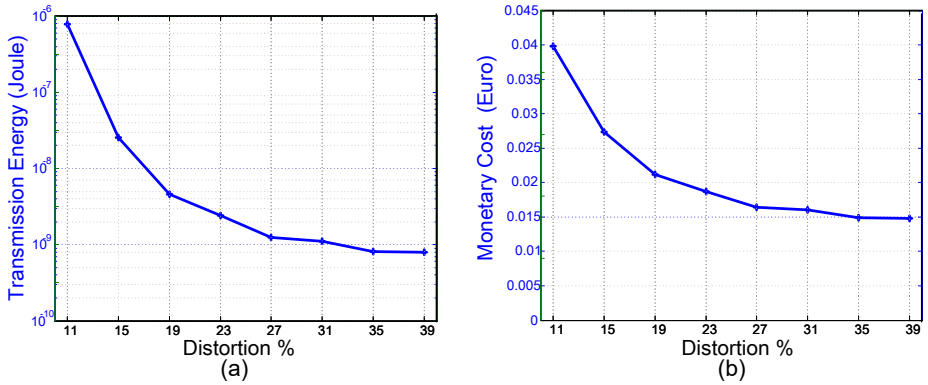


Fig. 4.7 Tradeoff between distortion and transmission energy in (a), and between distortion and monetary cost in (b), for $\lambda = 0.5$.

Next, we highlight the effect of λ on the value of the energy-cost-distortion tradeoff, i.e., on the value of the minimum U. At low λ, distortion is weighted more than transmission energy and monetary cost. Hence, the optimization problem results in low κ_i and low distortion, while the transmission energy and the monetary cost will be high, as shown in Figure 4.8-(a). Conversely, as λ increases (i.e., when the relevance of transmission energy and monetary cost increase) κ_i and the distortion level grow, as shown in Figure 4.8-(b)-(c). In these plots, the optimization problem is solved under the constraint that the distortion cannot exceed 30%, thus, as λ increases, the distortion grows until it reaches the maximum value.

These results stress that, when all performance metrics are equally important, it is paramount to adopt an algorithm, such as our proposed λ-DOA, which establishes the best tradeoff among transmission energy, monetary cost, and distortion. Indeed, at the optimal λ selected by the algorithm, we obtain the minimum value of transmission energy and monetary cost that allows satisfying the constraint on the maximum level of distortion.

4.6.3 Multiple PDA Scenario

Figure 4.9 depicts the performance of our distributed algorithm (Algorithm 3) compared to the centralized approach, and illustrates its convergence behavior to the centralized-optimal solution with varying τ. All PDAs are assumed to have the same delay deadline τ. It can be seen that the proposed algorithm converges in 60 iterations at most, however this number highly depends on the gradient step size β in

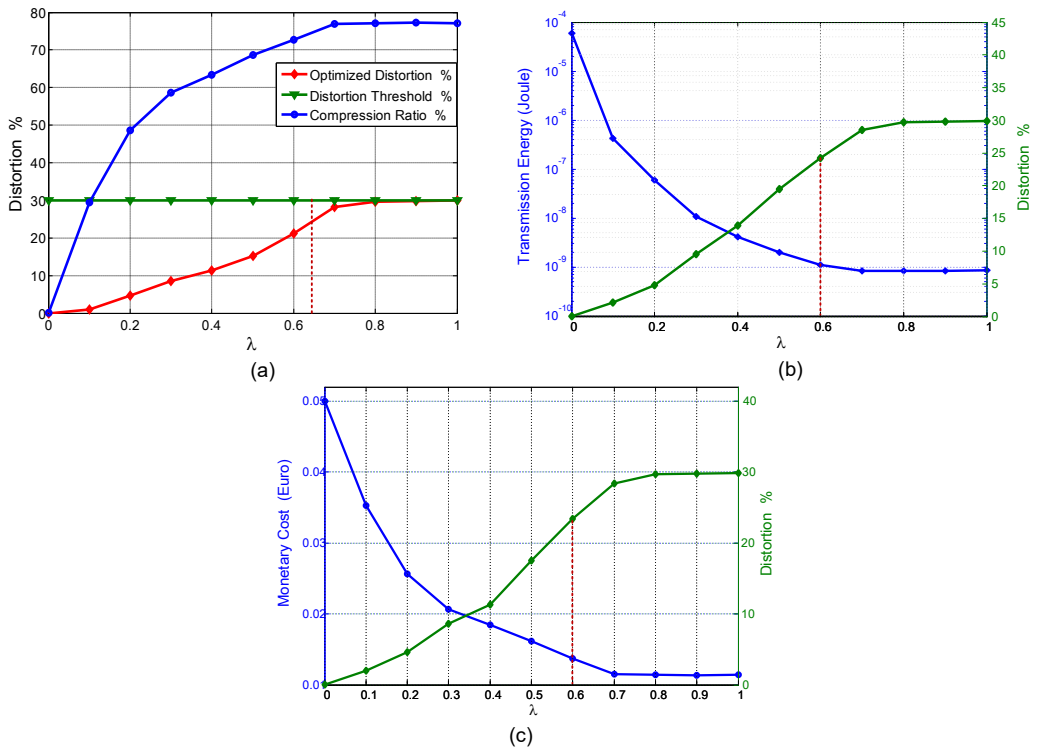


Fig. 4.8 Results for varying λ: (a) variation of compression ratio and distortion, (b) tradeoff between transmission energy and distortion, (c) tradeoff between cost and distortion.

(4.25). The optimal value of the objective function obtained through the distributed algorithm is compared to that of the centralized solution. The plot also shows the effect of varying τ on the optimal value of the objective function: by decreasing τ, each PDA increases its r_i to meet the delay deadline constraint. As a result, its transmission energy increases (see (4.2)). It follows that κ_i tends to increase so as to reduce the amount of transmitted data and achieve an optimal tradeoff between energy and distortion. This leads to an overall increase of the objective function.

In Figure 4.10, the performance of the proposed scheme is compared against a baseline algorithm in which the different PDAs evenly share the available bandwidth

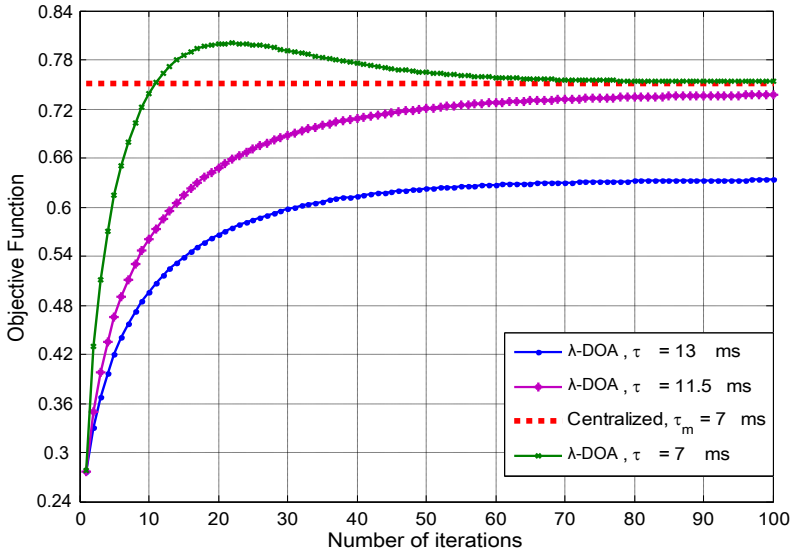


Fig. 4.9 Convergence of λ-DOA for different values of τ.

and each PDA i solves the following optimization problem:

$$\min_{\kappa_i, r_i} \; U_i(\kappa_i, r_i, \lambda) \quad \text{s.t.} \tag{4.26}$$

$$\frac{l_i(1-\kappa_i)}{r_i} \leq \tau \tag{4.27}$$

$$r_i \leq \frac{B}{N} \tag{4.28}$$

$$r_i \geq 0, \;\; 0 \leq \kappa_i \leq 1, \;\; 0 \leq \lambda \leq 1\,. \tag{4.29}$$

The problem solution at each PDA can be obtained using the same method as in Section 4.4. We will refer to this baseline scheme as Uniform Bandwidth Allocation (UBA). We remark that UBA imposes strict constraints on PDAs that have bad channel conditions, since it assigns equal bandwidth share to all PDAs. On the contrary, our scheme takes channel conditions into account. Thus, PDAs can transmit their data using variable bandwidth instead of being limited to fixed-assigned bandwidth, i.e., PDAs with bad channel conditions are assigned more bandwidth than others with good channel conditions, according to (4.14). This leads to a reduced energy consumption as well as distortion under λ-DOA compared to when UBA is used. Overall, λ-DOA offers about 15% improvement in the value of the objective function, with respect to the UBA scheme, as shown in Figure 4.10.

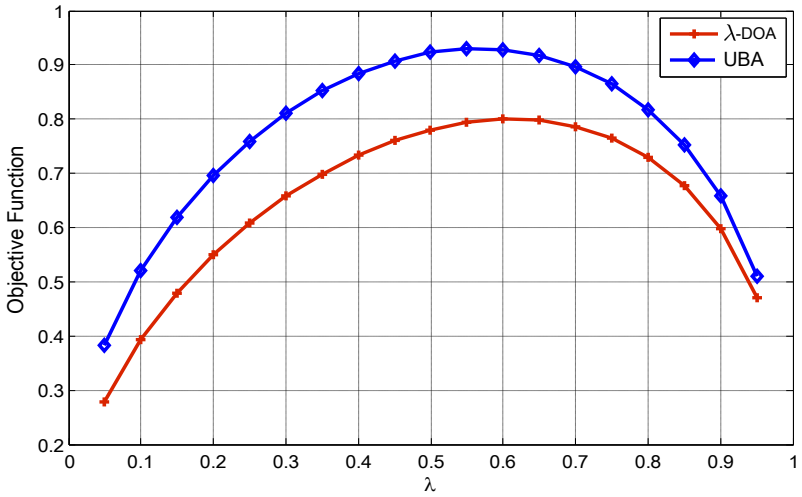


Fig. 4.10 Comparison between the value of the objective function obtained using the λ-DOA and the UBA scheme.

Next, we investigate the system performance when the maximum value of acceptable delay, τ, varies. Figures 4.11 and 4.12 depict the results in terms of total bandwidth usage (i.e., $\sum_{i=1}^{\tilde{N}} e^{\widehat{r_i}}$) as τ and λ change. We can observe that, as τ increases, the total bandwidth utilization decreases. This is due to the fact that, for a fixed value of distortion, the PDAs should reduce their transmission rates so as to minimize their energy consumption. When the effect of the weighting factor λ is considered, we note that the larger the λ, the higher the weight assigned to transmission energy and monetary cost at the expense of distortion. Consequently, lower r_i's are used and the difference in bandwidth usage for different values of τ results to be greatly magnified. This also illustrates the importance of λ as a tuning parameter that helps in fulfillment of network throughput constraint.

Figure 4.13 illustrates the value of the utility function U_i of three PDAs that share the same medium and send their data to the same MHC, as λ varies. The PDAs are assumed to have different channel propagation conditions and different distances from the MHC. Clearly, PDAs with bad channel conditions, such as PDA 3, are forced to increase their transmission energy in order to not exceed the maximum BER ϑ (see Figure 4.14). The higher energy consumption then leads to an increased distortion so as to maintain the desired energy-cost-distortion tradeoff. As a result, the utility function of the PDAs with harsh propagation conditions will be higher compared to that of PDAs with good channel (e.g., PDA 1)).

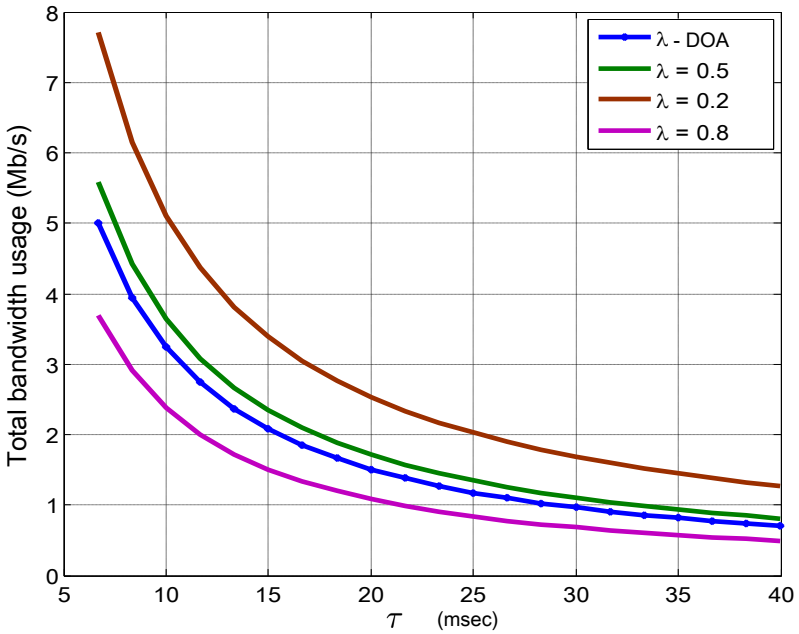


Fig. 4.11 Total bandwidth usage vs. τ, for different values of λ.

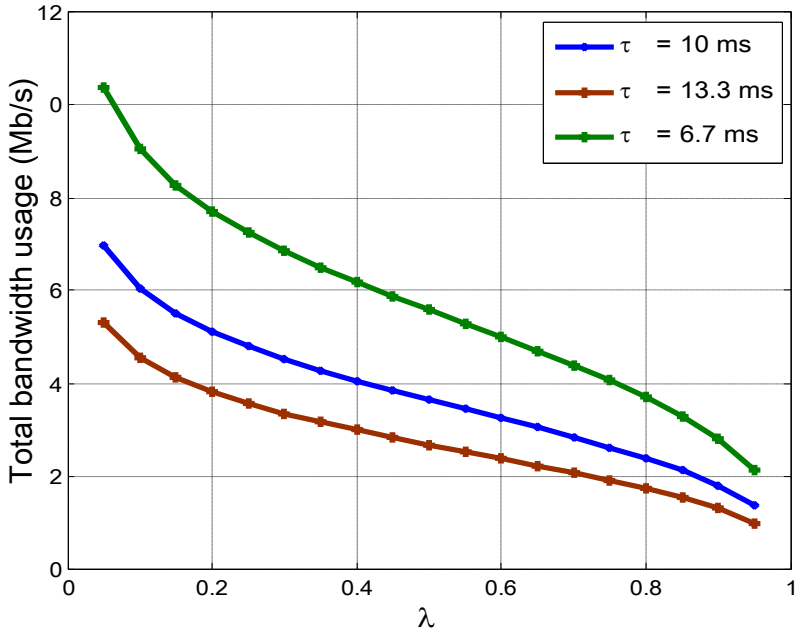


Fig. 4.12 Total bandwidth usage vs. λ, for different values of τ.

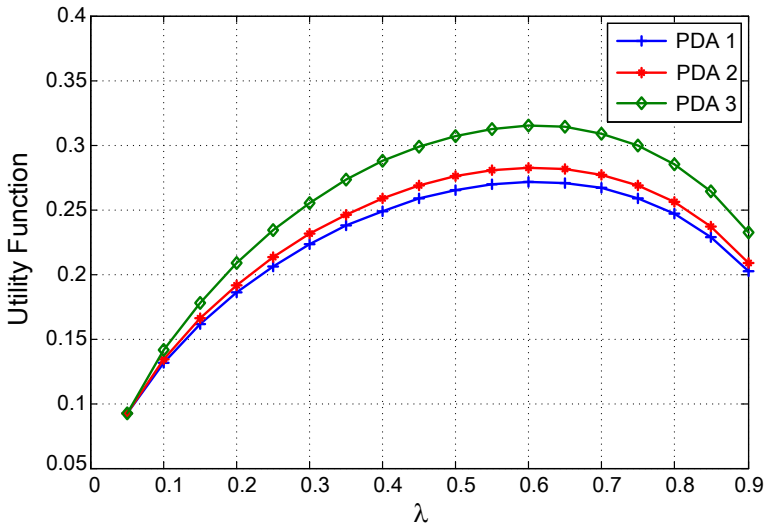


Fig. 4.13 Minimized utility function for three PDAs with different channel conditions (best: PDA 3, worst: PDA 1), λ varies.

Finally, in Figure 4.15 we assess the ability of our λ-DOA scheme to adapt to network dynamics. The impact of varying network load (i.e., number of admitted PDAs) and available bandwidth is studied. In Figure 4.15-(a), it is assumed that initially the network includes 3 PDAs. In this case, the λ-DOA algorithm converges to the optimal solution in around 50 iterations. After the first 100 time steps, one PDA leaves the network. As a result, the aggregate utility initially drops since the bandwidth that was allocated to the former PDA remains unused. However, thanks to the λ-DOA, the network resources are redistributed to the remaining PDAs and the network behavior converges again to the optimal solution. Note however that the aggregated utility is lower than in the case of 3 PDAs as the network is not saturated. At time step 200, a third PDA joins the network. The network quickly adapts to the new situation by assigning resources to the newly added PDA, leading to an increase in the aggregate utility.

In Figure 4.15-(b), again 3 PDAs initially participate in the network and the available bandwidth is set to $B = 4$ Mbps. As before, the λ-DOA scheme converges to the optimal solution. At time step 100, the available bandwidth is doubled to $B = 8$ Mbps. The network adapts to the change in the available bandwidth by allocating more bandwidth to the PDAs. Note that, due to the larger available bandwidth, the PDAs increase their transmission rates and decrease their compression ratio, thus achieving a significant reduction in the distortion level. It follows that the

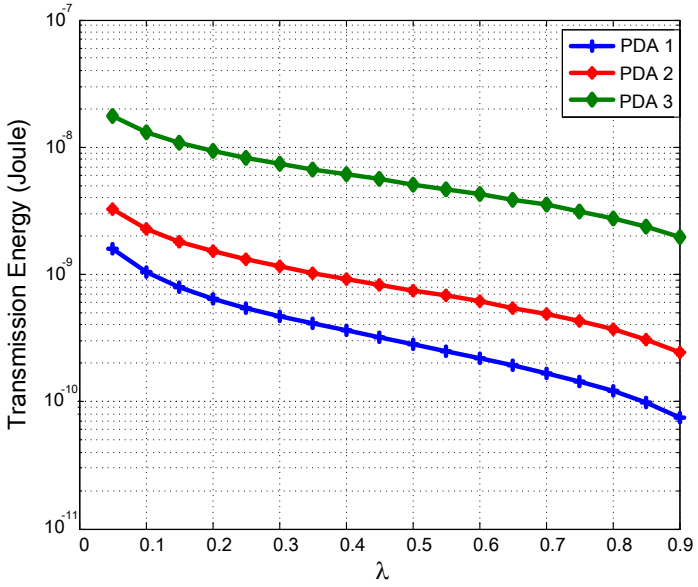


Fig. 4.14 Transmission energy consumed by three PDAs with different channel conditions (best: PDA 1, worst: PDA 3), λ varies.

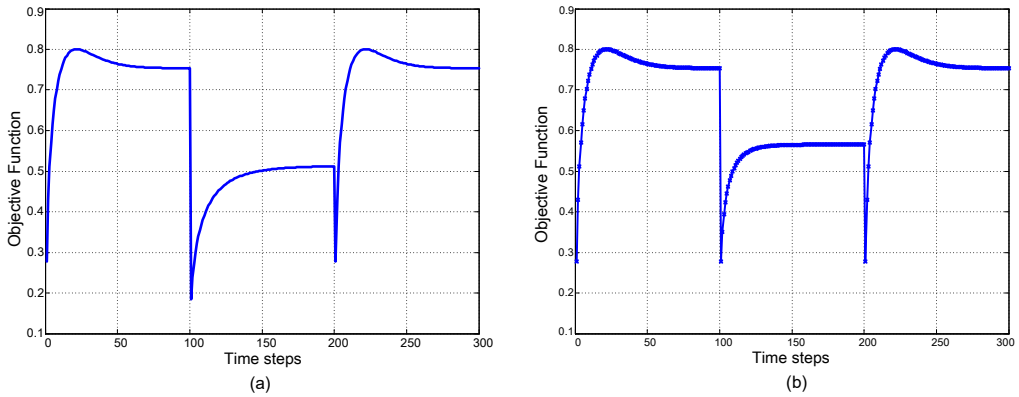


Fig. 4.15 Temporal evolution of the system performance with varying (a) number of participating PDAs and (b) available bandwidth.

new optimal solution corresponds to a lower value of the objective function. After that, the available bandwidth drops again to $B = 4$ Mbps. The network dynamically adapts to the new situation, converging to the initial solution.

4.7 Summary

This chapter addressed the problem of optimizing the transmission of m-health applications data from energy-constrained PDAs to the MHC, which is in charge of reconstructing medical signals, such as EEG, with low distortion. We therefore take transmission energy, monetary cost, and signal distortion as main performance metrics, while accounting for the radio propagation conditions experienced by the PDAs. We proposed a multi-objectives optimization problem that aims at establishing the desired tradeoff among our main performance metrics, while meeting the system constraints in terms of maximum data transfer latency and BER. The centralized multi-objective resource optimization problem has been decomposed into a set of sub-problems that can be solved in a distributed, efficient manner. According to the proposed algorithm, the PDAs can separately calculate their data transfer parameters (i.e., compression ratio and transmission rate) through the exchange of a limited number of control messages with the MHC. Furthermore, we proposed an algorithm that allows to achieve the optimal tradeoff among the main performance metrics when all of them are equally relevant to the system. Simulation results demonstrate that the proposed scheme obtains the optimal tradeoff between energy efficiency and QoS requirements, while providing 15% savings in the objective function (i.e., energy-cost-distortion utility function), compared to solutions based on equal bandwidth allocation.

Chapter 5

Dynamic Networks Association with Adaptive Data Compression for Smart Health systems

5.1 Overview

Rapid advances in wireless access technologies and in-network processing can significantly assist in implementing smart healthcare systems through providing seamless integration of heterogeneous wireless networks, medical devices, and ubiquitous access to data. S-health demand for high data rates and QoS has motivated us to leverage the development of cellular networks into dense heterogeneous networks (HetNets) with the utilization of multi-Radio Access Technology (RAT). It is essential for each user/device to leverage different RATs, hence, the available radio resources across different spectral bands, to communicate with the network infrastructure [135]. Utilization of the spectrum across diverse radio technologies is expected to significantly enhance network capacity and QoS for emerging applications such as remote healthcare monitoring. However, this imposes an essential need to develop innovative networks association mechanisms that account for energy efficiency while meeting application quality requirements.

In accordance with the new trends foreseen for 5G systems, this chapter proposes an efficient networks association mechanism with adaptive data compression for en-

hancing the performance of s-health systems.[1] Different performance matrices have been considered, in addition to networks characteristics and application requirements, in order to find an efficient solution that grasps the conflicting nature of the various users' objectives and addresses their inherent tradeoffs. The proposed mechanism adopts a user-centric approach towards exploiting heterogeneous wireless networks to optimize medical data delivery over heterogeneous s-health systems. In particular, we focus on answering the following questions:

1. Which network(s) should be selected among multiple Radio Access Networks (RANs)?
2. What is the optimal level of data compression to be used?
3. What is the amount of data that should be sent through each selected RAN after compression?

While addressing the above issues, we account for both network characteristics and application requirements, providing a solution, which achieves an optimal energy-quality-cost tradeoff.

5.2 Related Work

With the rising tendency toward network densification, various radio technologies, such as 3G, 4G and WiFi, could be jointly leveraged to enable seamless connection to users with high levels of quality of experience [137]. Simultaneously utilizing of multiple radio technologies turns to be even more serious as the user demand and QoS requirements proliferate, while the available wireless resources remain limited. Consequently, the upcoming 5G systems are expected to have dense and irregular HetNets, where the user will be able to access the system through different points of access. Thus, it is crucial to develop techniques that can efficiently leverage the available radio resources across different spectral bands using multi-RAT [108].

The association with network infrastructure may be concurrent, exploiting the multihoming feature of mobile devices to establish simultaneous associations with different access networks, or switch from one point of access point to another, within

[1] This work has been published in [136].

the same RAN or across different RANs. In both cases, several schemes have been proposed in the literature for network selection and association in HetNets. The proposed approaches can be broadly classified into four categories: cost-function based, decision making processes using game theory, Markov decision processes (MDPs), and optimization based. Cost function-based schemes proactively select the network with the highest/lowest utility or cost function [127][138]. Although the approach can achieve a near optimal solution, it is often hard to prove it. According to the game theory approach, instead, users in different service areas compete for the bandwidth offered by different wireless networks [139][140]. The resulting algorithms are, in general, complexity-prohibitive, and their convergence is not guaranteed. Even in case of convergence, they do not necessarily converge to an optimal solution. MDPs have been used also to study network switching between different RATs [141][142]. However, finding the optimal solutions is again cumbersome, especially in the case of large networks [137]. Formulating network selection problem as an optimization problem with low or moderate complexity is also not a trivial task. Finding optimal resource allocation and user association, subject to resources and/or power constraints, may result in an NP-hard problem [143]. One way to make the problem tractable is by using constraints relaxation or variables transformation, or by envisioning online adaptive methods such as Q-learning [144].

Finally, the existing work on concurrent association mainly focus on designing a traffic scheduler over different device interfaces considering users incentives for collaboration and bandwidth sharing [145], a transport-layer control protocol to enable concurrent multipath transport [146], or content-aware transport-layer protocols [147]. Other studies have focused on the resource allocation problem for parallel transmission considering multi-RATs [148][149]. For instance, a sub-optimal solution is presented in [148] by utilizing the intrinsic quasi-concavity of the formulated problem. While, in [149], the authors develop a framework of multi-RAT system, where a small cell serves a number of mobile users via IEEE 802.11 WLAN and 3GPP LTE access technologies. A scheduler at the small cell is developed in order to minimize the total transmission power subject to QoS constraints on the users' transmission rates. In [150], an urban deployment scenario is investigated, where WiFi small cells are overlaid on top of the 3GPP LTE network. The authors present user-centric network selection algorithms to minimize feedback overhead while considering user's preferences. A good review of the mathematical methods that are applied to the network selection problem, including cost-function, game theory,

multiple attribute decision making, combinatorial optimization, fuzzy logic, and Markov chain, can be found in [151].

To the best of our knowledge, none of the aforementioned work advocates a user-centric approach for efficient networks association taking into consideration context-aware in-network processing to optimize the delivery, cost, and latency of the medical data. Thus, our main contributions are summed up as follows:

1. A multi-objective optimization problem is formulated that enables each PDA to optimally set its data compression ratio and select the RAN(s) for data transmission, in an energy-efficient and cost-aware manner while ensuring an acceptable signal distortion.

2. We propose an analytical solution for the formulated optimization problem, by decomposing it into two sub-optimizations. The two sub-problems turn out to be solvable with low complexity, and they are analytically proved to lead to the same optimal solution as the original problem.

3. We design a distributed, iterative, PDA-centric algorithm considering the case where initially PDAs have just a rough estimate of their resource share on the available RANs. Then, the convergence behavior of the proposed algorithm to the optimal solution is analyzed and proved analytically.

4. Finally, the performance of the proposed approach is evaluated and compared against that of state-of-the-art techniques. Our results demonstrate that the proposed solution allows for high-quality healthcare monitoring of patients, it significantly outperforms other solutions, and swiftly adapts to varying network dynamics.

5.3 System Model and Performance Metrics

Implementing s-health systems in heterogeneous multi-radio environment is challenging. We consider the s-health architecture shown in Figure 5.1, which is divided into three main sectors: data acquisition and pre-processing, wireless multi-RAT network, and health monitoring services and applications. The first sector represents patients equipped with BASN, pre-attached to their PDAs. Herein, the smartphones can act as PDAs, i.e., they collect the medical data from the sensors/medical devises

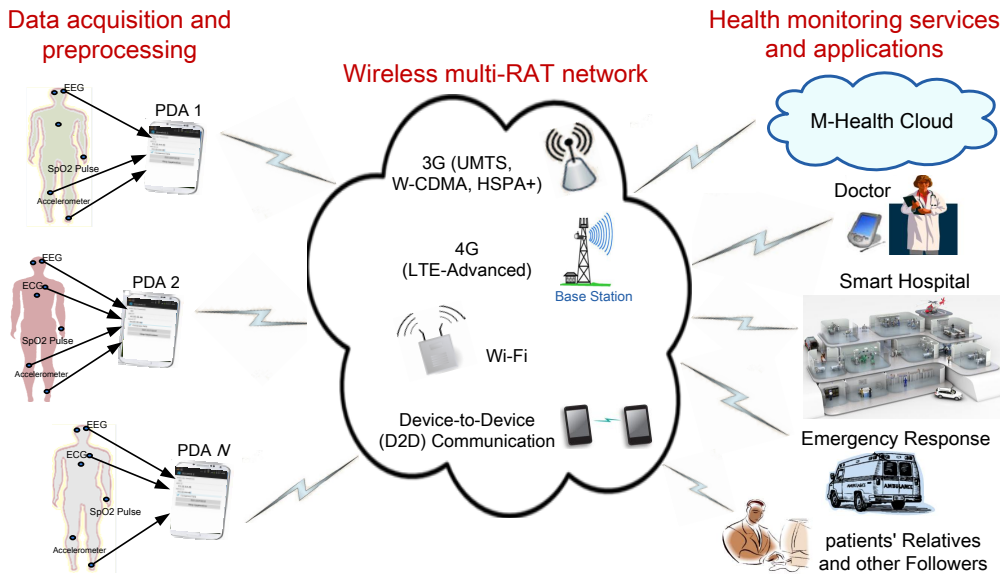


Fig. 5.1 Multi-RAT m-Health system scenario.

and execute in-network processing tasks to optimize the data transfer based on the context and the network conditions. More specifically, the PDA may compress the gathered data at the cost of a certain degree of signal distortion, and transfer it to the MHC over the multi-RAT wireless network. Importantly, the available multi-RAT environment enables the PDA to be connected anytime and anywhere, given that innovative networks selection schemes are implemented. Furthermore, it is of prominent importance that the data transfer from the PDAs to the MHC takes place in an energy-efficient manner, in order to ensure a long lifetime of the battery-operated PDAs.

Although the proposed framework considers the encoding model of EEG signals [120], it can be easily extended to diverse biosignals and multimedia data. In the following, we consider a time period T assuming that each PDA i ($i = 1, \ldots N$) has to transfer B_i bits of data toward the MHC. As mentioned, each PDA compresses the gathered data so that the actual amount of bits to be transferred is given by: $b_i = B_i(1 - \kappa_i)$ with κ_i being the data compression ratio used by PDA i. As a result, a signal distortion is introduced due to the data compression, which is expressed in (4.1).

Now, assume that RAN j operates on a bandwidth W_j and that the generic PDA i has a data rate r_{ij} on RAN j. Clearly, r_{ij} relies on the access technology (e.g., its maximum value is 54Mbps in IEEE 802.11a/g) and on channel radio propagation

conditions. As noted in [120], the estimated energy consumption for PDA i to transfer b_i bits over RAN j is:

$$\widetilde{E}_{ij} = \psi_j \left(\frac{b_i N_0 W_j}{r_{ij} g_{ij}} (2^{\frac{r_{ij}}{W_j}} - 1) \right) + c_j \,. \tag{5.1}$$

In the above expression, N_0 is the noise spectral density, while the channel gain g_{ij} is defined as

$$g_{ij} = K \cdot \sigma \cdot |h_{ij}|^2$$

where $K = -1.5/(\log(5\text{BER}))$, σ is the path loss attenuation, and $|h_{ij}|$ is the fading channel magnitude for PDA i over RAN j. All related parameters for this model are defined as in [120]. In (5.1), ψ_j and c_j are specific parameters that differ for each network interface [122]. They can be acquired from the radio interface specifications, or calculated through power consumption measurements [127].

Next, looking at the expected latency provided by each RAN, we define:

$$\widetilde{L}_{ij} = \frac{b_i}{r_{ij}} + \xi_j, \tag{5.2}$$

where $\frac{b_i}{r_{ij}}$ and ξ_j are, respectively, the air time and the access channel delay that PDA i expects to experience when transmitting b_i bits through RAN j. In other words, it represents the estimated end-to-end delay when using a given technology [152].

Finally, the monetary cost (hereinafter referred to as cost for brevity) resulting from using RAN j by PDA i to send b_i bits is expressed in Euro and defined as:

$$\widetilde{C}_{ij} = b_i \varepsilon_j \tag{5.3}$$

where ε_j is the monetary cost per bit for RAN j. This cost can be obtained through the use of, e.g., the IEEE 802.21 standard [153], which allows a user device to collect information about the available wireless networks [127]. Such value can also be stored on the PDAs in advance and updated if there are any changes in pricing.

5.4 Joint Network Selection and Compression Optimization: Problem Formulation and Solution

Looking at the system model and the performance provided by each RAN (5.1)–(5.3), it can be seen that there is a tradeoff between distortion on one side and energy consumption, latency and cost on the other. First, the higher the compression ratio (κ_i), the greater the distortion, but the smaller the amount of data to transmit (b_i). Secondly, as the data rate over RAN j (r_{ij}) increases, the energy consumption increases, while the latency decreases. Also, it is often the case that RANs providing higher data rates and lower latency have a higher monetary cost.

In a system supporting healthcare applications, it is of paramount importance to provide an acceptable distortion level and to ensure a swift transfer of medical data towards the MHC. Then, from a practical point of view, it is crucial that PDAs do not have to be recharged too often and that services have an acceptable cost. Thus, in light of such requirements, in Section 5.4.1 we formulate a Multi-objective Optimization Problem (MOP) that each PDA should solve and whose aim is to find the optimal tradeoff between the above conflicting objectives. The proposed problem is then analytically solved in Section 5.4.2, where the proposed problem is decomposed into two sub-optimization problems: network selection optimization, and compression optimization. In Section 5.4.3, we present an iterative algorithm to solve the network selection problem, while in Section 5.4.4, we derive a closed-form solution for the compression optimization. Finally, a practical algorithm is introduced in Section 5.5 for adaptive network selection and compression, which applies to the case where only an initial estimate of the RANs' resources is available at the PDAs.

5.4.1 Problem Formulation

The objective of the proposed MOP is threefold: (i) minimizing transmission energy consumption, (ii) minimizing monetary cost, and (iii) meeting the medical data QoS requirements in terms of signal distortion and data delivery latency. We therefore define a single aggregate objective function which turns the above multiple objectives into a single objective function. However, each objective presents different ranges and units of measurement, hence we first normalize these quantities with respect

to their maximum value, in order to make them adimensional and comparable. We will denote the normalized energy, monetary cost, and latency by E_{ij}, C_{ij} and L_{ij}, respectively.

Given a generic PDA i with M available RANs, the objective of our optimization problem is to obtain the optimal compression ratio and assign the PDA to the optimal RAN(s) minimizing the transmission energy consumption E_{ij}, monetary cost C_{ij}, latency L_{ij}, and distortion D_i:

$$\textbf{P:} \quad \min_{P_{ij},\kappa_i} \sum_{j=1}^{M} P_{ij}U_{ij} + \delta_i D_i \tag{5.4}$$

$$\text{s.t.} \quad \frac{P_{ij}\cdot b_i}{r_{ij}} \leq T_{ij}, \ \ \forall j \in M \tag{5.5}$$

$$\sum_{j=1}^{M} P_{ij} \geq 1, \tag{5.6}$$

$$0 \leq P_{ij} \leq 1, \quad \forall j \in M \tag{5.7}$$

$$0 \leq \kappa_i \leq 1. \tag{5.8}$$

where $U_{ij} = \alpha_i E_{ij} + \beta_i C_{ij} + \gamma_i L_{ij}$ is the utility function of PDA i over RAN j. The weighting coefficients represent the relative importance of the four objective functions in the problem; it is assumed that $\alpha_i + \beta_i + \gamma_i + \delta_i = 1$.[2] Moreover, in (5.4) we consider a network utilization indicator P_{ij} that represents the fraction of data that should be transmitted through RAN j by PDA i. Note that PDAs have all information to compute the expected energy consumption and cost. Additionally, we assume that the RAN notifies the PDA about the physical data rate r_{ij}. The network also notifies PDAs about the expected channel access delay (ξ_j) the PDA may experience.

The network capacity constraint is represented by (5.5), where T_{ij} is the maximum fraction of the time period T that can be used by PDA i over RAN j (resource share). T_{ij} depends on the number of PDAs accessing the RAN, and we assume that it is notified by the RAN. Constraint (5.6) instead ensures that all the data that PDA i has to transfer to the MHC is actually sent through the wireless medium.

[2] In [154], a dynamic weights update mechanism is proposed in order to achieve the fairness among different user's objectives (i.e., energy saving, monetary cost and service latency), while enhancing the devices operating time.

The unknowns in this problem are the P_{ij}'s and κ_i, i.e., each PDA needs to determine its compression ratio and the amount of data that the PDA should transfer through the different RANs. Looking at problem formulation in (5.4), one can see that it is not a linear programming (LP) problem [128], due to the terms involving the product of P_{ij} by κ_i (or functions of κ_i). Also, a simple approach, like transforming the problem into a Geometric Program (GP), would not work in this case due to the existence of the constraint in (5.6) with the non-linearity of the distortion objective. Thus, below we envision a methodology to decompose the problem into two sub-optimization problems, for which an optimal, analytical solution can be obtained.

5.4.2 Optimization Decomposition

In order to analytically solve (5.4), one would like to break the original problem into two sub-problems such that each of them is a function of one decision variable only and, hence, can be solved independently of the other. The difficult point in our case is that the optimization variables (i.e., P_{ij}'s and κ_i) are coupled. To overcome this issue, we proceed as follows.

We first look at the optimization variables in (5.4) as network selection variables P_{ij}'s and adaptive compression variables κ_i. Network selection variables can be considered as global variables that are relevant to the overall system, while adaptive compression variables are local variables at each PDA. We therefore decompose the problem into the network selection and adaptive compression sub-problems, and we prove that solving the new problem formulation still leads to the optimal solution of the original problem in (5.4).

Theorem 1**:** *The optimization problem in (5.4) can be decomposed into two sub-optimization problems while maintaining the optimal solution, as follows:*

$$\textbf{SP1:} \quad \min_{P_{ij}} \sum_{j=1}^{M} P_{ij}\hat{U}_{ij} \tag{5.9}$$

$$\text{s.t.} \quad \frac{P_{ij}b_i}{r_{ij}} \leq T_{ij}, \ \ \forall j \in M \tag{5.10}$$

$$\sum_{j=1}^{M} P_{ij} \geq 1, \tag{5.11}$$

$$0 \leq P_{ij} \leq 1, \quad \forall j \in M \tag{5.12}$$

and

$$\textbf{SP2:} \quad \min_{\kappa_i} \left(\delta_i D_i - \sum_{j=1}^{M} P_{ij} \kappa_i \bar{U}_{ij} \right) \tag{5.13}$$

$$0 \leq \kappa_i \leq 1 \tag{5.14}$$

where $\bar{U}_{ij}$ and $\hat{U}_{ij}$ have a similar expression as U_{ij} but for some constant terms and the fact that b_i is replaced by B_i, i.e., they are independent of κ_i.

Proof. The distortion term D_i (see Eq. (4.1)) is not a function of P_{ij}. Thus, the objective function in (5.4) can be written as

$$\min_{P_{ij},\kappa_i} \sum_{j=1}^{M} P_{ij} U_{ij} + \min_{\kappa_i} \ \delta_i D_i \,. \tag{5.15}$$

By denoting with E_M, C_M and L_M, the maximum energy expenditure, cost, and latency, respectively, we can rewrite Eqs. (5.1), (5.2), and (5.3), as:

$$\begin{aligned}
E_{ij} &= \frac{\left[(1-\kappa_i)\psi_j \left(\frac{B_i N_0 w_j}{r_{ij} g_{ii}} (2^{r_{ij}/w_j} - 1) \right) + c_j \right]}{E_M} \\
&= (1-\kappa_i)\bar{E}_{ij} \\
C_{ij} &= \frac{(1-\kappa_i) B_i \varepsilon_j)}{C_M} = (1-\kappa_i)\bar{C}_{ij} \\
L_{ij} &= \frac{(1-\kappa_i) B_i}{r_{ij} L_M} + \frac{\xi_j}{L_M} = (1-\kappa_i)\bar{L}_{ij} + \hat{L}_{ij}
\end{aligned} \tag{5.16}$$

where $\bar{E}_{ij}$, $\bar{C}_{ij}$, $\bar{L}_{ij}$, and $\hat{L}_{ij}$ are independent of κ_i (as well as P_{ij}). By substituting (5.16) in (5.15), we get:

$$
\begin{aligned}
Z &= \min_{P_{ij},\kappa_i} \sum_{j=1}^{M} P_{ij}\left(\alpha_i \bar{E}_{ij} + \beta_i \bar{C}_{ij} + \gamma_i(\bar{L}_{ij} + \hat{L}_{ij})\right) \\
&\quad - \sum_{j=1}^{M} P_{ij}\kappa_i(\alpha_i \bar{E}_{ij} + \beta_i \bar{C}_{ij} + \gamma_i \bar{L}_{ij}) \\
&\overset{(a)}{=} \min_{P_{ij}} \sum_{j=1}^{M} P_{ij}\hat{U}_{ij} - \min_{P_{ij},\kappa_i} \sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij} \\
&= \min_{P_{ij}} \sum_{j=1}^{M} P_{ij}\hat{U}_{ij} + \max_{P_{ij},\kappa_i} \sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij}
\end{aligned} \tag{5.17}
$$

where in *(a)* $\hat{U}_{ij} = \bar{E}_{ij} + \bar{C}_{ij} + (\bar{L}_{ij} + \hat{L}_{ij})$ and $\bar{U}_{ij} = \bar{E}_{ij} + \bar{C}_{ij} + \bar{L}_{ij}$.

Now, to minimize Z, we need to minimize $\sum_{j=1}^{M} P_{ij}\hat{U}_{ij}$ and at the same time maximize $\sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij}$. However, to maximize $\sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij}$, we need to maximize $P_{ij}\bar{U}_{ij}$, where $\hat{U}_{ij}$ differs from $\bar{U}_{ij}$ by an additive positive constant. Thus, this is conflict with the minimization of $\sum_{j=1}^{M} P_{ij}\bar{U}_{ij}$. Thus, the only possible solution is to minimize $\sum_{j=1}^{M} P_{ij}\hat{U}_{ij}$ with respect to P_{ij} and maximize $\sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij}$ with respect to κ_i. Accordingly, we will have:

$$
\begin{aligned}
Z &= \min_{P_{ij}} \sum_{j=1}^{M} P_{ij}\hat{U}_{ij} + \max_{\kappa_i} \sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij} \\
&= \min_{P_{ij}} \sum_{j=1}^{M} P_{ij}\hat{U}_{ij} - \min_{\kappa_i} \sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij}.
\end{aligned} \tag{5.18}
$$

This proves that the original optimization problem in (5.4) can be decomposed into two sub-optimization problems, (5.9) and (5.13), which still lead to the optimal solution. □

It is worth mentioning that decoupling the overall optimization problem into two sub-problems, greatly simplifies the problem, thus allowing the study of different adaptive compression techniques with different distortion models. Similarly, the usage of various video coding schemes for medical video content delivery could be investigated, in the presence of different network types and network conditions.

5.4.3 Network Selection Optimization

In the following, we present an analytical solution for the network selection optimization problem in (5.9). The problem is a LP problem. Thus, we can reduce the objective function by increasing P_{ij}'s with minimum $\hat{U}_{ij}$. Since these variables have non-negative coefficients, there would be no other way to decrease the objective function. We conclude that in the following proposition.

Proposition 1**:** *The optimal solution of (5.9) can be obtained by maximizing the values of P_{ij}'s for which the corresponding $\hat{U}_{ij}$'s are minimum.*

Proof. Initially, assume that $P^* = 0$ is the optimal feasible solution, where P^* is a vector of the variables P_{ij}'s, with corresponding objective value $Z^* = 0$. However, the values of P_{ij}'s have to be increased in order to satisfy the constraint in (5.5), i.e., the optimal solution must be $P^* > 0$. Next, recall that $\hat{U}_{ij}$ depends on the characteristics of RAN j (e.g., data rate, bandwidth, cost per bit, etc.). By contradiction, we can show that the optimum solution P^* is the one for which the values $\hat{U}_{ij}$'s are minimum, i.e., $P^* = \tilde{P}$ such that $\tilde{U}_{ij} = \min \hat{U}_{ij}\ \forall j$. Indeed, consider that $P^* = \tilde{P} + \Delta$ instead. In this case, $\hat{U}(P^*) > \hat{U}(\tilde{P})$ with $\hat{U}(P)$ being the vector of the values $\hat{U}_{ij}$'s obtained for P, and $P^*\hat{U}(P^*) > \tilde{P}\hat{U}(\tilde{P})$. This contradicts the assumption that $\tilde{P} + \Delta$ is the optimal solution, hence the only optimal solution is $P^* = \tilde{P}$; any other feasible solution will have a strictly larger objective value. □

According to the above proposition, we can solve the network selection optimization problem in (5.9) using Algorithm 5, while maintaining the optimal solution. The algorithm sorts the available RANs in ascending order, according to the values $\hat{U}_{ij}$'s, then the network with the lowest $\hat{U}_{ij}$ is selected, and P_{ij} is calculated as:

$$P_{ij} = \max\left(1, \frac{T_{ij} r_{ij}}{b_i}\right). \tag{5.19}$$

Note, indeed, that P_{ij} cannot exceed 1, according to the constraint in (5.7). If the available resource share on this RAN does not satisfy the requirement of the PDA (i.e., $\sum_{j=1}^{M} P_{ij} < 1$), the second RAN in the list is selected, and so on till the constraint in (5.6) is satisfied. The algorithm complexity is closely related to the number of available RANs M; the worst-case complexity is $O(M \log(M))$.

Algorithm 5 Network Selection Optimization at PDA i

Require: α_i, β_i, γ_i, T_{ij}, r_{ij}, ε_j

```
p = 1.
Sort the available RANs according to the Û_ij values in ascending order
for j = 1 → M do
    Compute P_ij according to (5.19)
    if P_ij ≥ p then
        Set P_ij = p, and P_ik = 0, ∀k s.t. j < k ≥ M.
        Break % Constraint (5.6) is met
    else
        p = p − P_ij
    end if
end for
return Selected RAN(s) and corresponding optimal P_ij's
```

5.4.4 Adaptive Compression Optimization

As far as the problem in (5.13) is concerned, a closed-form expression for the solution can be obtained by imposing that the derivative with respect to κ_i of the objective function is equal to 0. I.e.,

$$\begin{aligned} \partial/\partial\kappa_i &= \partial/\partial\kappa_i \left[\delta_i D_i - \sum_{j=1}^{M} P_{ij}\kappa_i \bar{U}_{ij} \right] \\ &= \delta_i \frac{\partial D_i}{\partial \kappa_i} - \sum_{j=1}^{M} P_{ij}\bar{U}_{ij} = 0. \end{aligned} \tag{5.20}$$

The distortion in (4.1) can be approximated[3] as,

$$D_i \approx \frac{c_2(1-\kappa_i)^{-c_3} + c_4 F^{-c_5} - 6.5}{100}. \tag{5.21}$$

[3]This simplified expression is still extremely accurate with a mean square error that equals 0.1%, while enabling us to maintain a closed-form expression of the solution (see Figure 5.2).

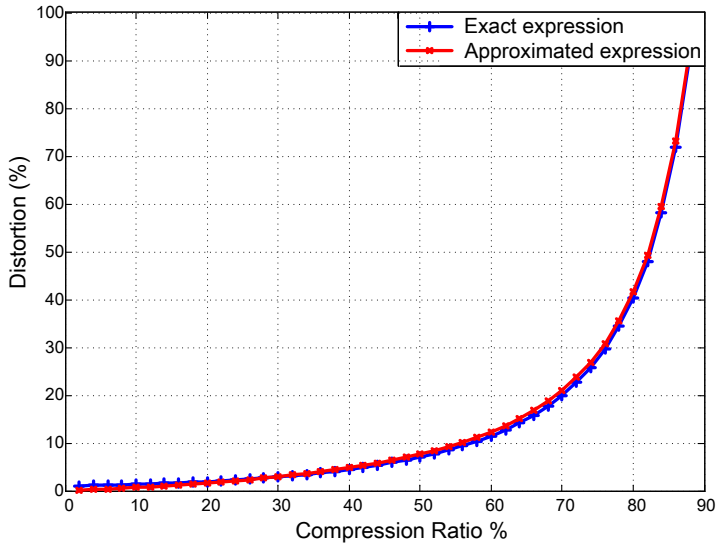


Fig. 5.2 Comparison between exact distortion obtained through the expression in (4.1) and the approximated value computed through (5.21).

By substituting (5.21) in (5.20), we obtain:

$$\begin{aligned}
\frac{\partial D_i}{\partial \kappa_i} &= \frac{\sum_{j=1}^{M} P_{ij}\bar{U}_{ij}}{\delta_i} \\
\frac{c_2 c_3 (1-\kappa_i)^{-c_3-1}}{100} &= \frac{\sum_{j=1}^{M} P_{ij}\cdot\bar{U}_{ij}}{\delta_i} \\
(1-\kappa_i)^{-c_3-1} &= \frac{100\sum_{j=1}^{M} P_{ij}\cdot\bar{U}_{ij}}{\delta_i\cdot c_2\cdot c_3} \\
\log(1-\kappa_i) &= -\frac{\log\zeta}{(1+c_3)}
\end{aligned} \tag{5.22}$$

where

$$\zeta = \frac{100\sum_{j=1}^{M} P_{ij}\cdot\bar{U}_{ij}}{\delta_i c_2 c_3}.$$

Thus, according to (5.22), the optimal κ_i is given by:

$$\kappa_i = 1 - \zeta^{-\frac{1}{1+c_3}}. \tag{5.23}$$

5.5 Adaptive Network Selection and Compression

In this section, we propose a distributed, iterative algorithm for optimal Adaptive Network Selection and Compression, named ANSC for short. ANSC leverages the problem decomposition introduced in the previous section, and it aims at finding the optimal solution of (5.4) in practical scenarios where PDAs may have just an initial estimate of their resource share on a given RAN j.

According to ANSC, once obtained the list of the available RANs, each PDA i initially assumes that no compression is performed (i.e., $\kappa_i = 0$) and runs Algorithm 5 locally, in order to find the optimal values of P_{ij} that determine which network(s) i should use and the amount of data that i should transfer on each RAN. Recall that in Algorithm 5 the weights α_i, β_i and γ_i are assumed to be pre-defined according to application requirements and/or PDAs' preferences. Importantly, the value T_{ij} is initially set to $T_{ij} = T_j/N_j, \forall j$, where N_j is the number of PDAs using RAN j, i.e., all PDAs assume to receive the same resource share on RAN j.[4] As foreseen by several standards, the RAN can notify users about the value of N_j.

Next, the generic PDA obtains the optimal κ_i using Eq. (5.23). It then broadcasts the corresponding value of $\widetilde{T}_{ij}$'s, i.e., the amount of resources it intends to "consume" over RAN j, which is given by:

$$\widetilde{T}_{ij} = \frac{P_{ij} b_i}{r_{ij}}. \tag{5.24}$$

At the RAN point of access, the actual demand from all PDAs is calculated, and each RAN j can use whatever mechanism to allocate the remaining resources among competing users (e.g., using proportional fairness, round robin, etc.) [155]. Each RAN can then return to the PDAs the values of their actual resource share. Accordingly, the PDAs run network selection optimization (Algorithm 5) again, obtaining the updated optimal values of P_{ij}'s. The procedure can be repeated until the convergence or a maximum number of iterations have been reached. The main steps of the ANSC algorithm are illustrated in Algorithm 6.

Below, we prove that the convergence of ANSC scheme is guaranteed.

[4]Note that the value of T_{ij} can be initially set to any arbitrary value.

Algorithm 6 Adaptive Network Selection and Compression (ANSC) algorithm at PDA i

Initialization: $\kappa_i = 0, t = 0$
$S(t) = 0$
do
Get optimal P_{ij}'s through Algorithm 5
Compute κ_i using (5.23)
Broadcast requested $\widetilde{T}_{ij}$'s computed through (5.24)
Get updated $T_{ij}(t+1)$ from RAN j
$t++$
$S(t) \leftarrow$ Value of objective function in (5.4)
while $|S(t) - S(t-1)| > \varepsilon \wedge t < n_{iter}$
return Selected RAN(s); optimal P_{ij}'s and κ_i

Theorem 2**:** *Regardless of the scheduling mechanisms implemented at the available RANs, the ANSC scheme converges to the optimal solution of the optimization problem in (5.4).*

Proof. The ANSC algorithm initially starts assuming an equal resource share among PDAs and no compression $\kappa_i = 0$. Since instead compression can be used, the actual resource share consumed by any PDA will be less than or equal to this initial arbitrary value, as long as no new users join the RANs (i.e., N_j is fixed). In other words, after obtaining the optimal κ_i using (5.23), the length of the transmitted data, hence the actual amount of radio resources consumed by the PDAs, decreases. As a result, there will be some extra share of resources on certain RANs that were not available at the previous iteration. No matter the scheduling mechanism used at the RANs to reallocate the free resources, for any arbitrary PDA i, we will have:

$$T_{ij}(t+1) \geq T_{ij}(t), \quad \forall t, j. \tag{5.25}$$

As a result, at each iteration the constraint in (5.5) becomes looser and looser, and the objective function can be decreased by increasing the value of the P_{ij}'s corresponding to the lowest U_{ij}'s. Thus, from (5.25) we can conclude that the objective function will always decay as the number of iterations increases until convergence is reached, and this will happen regardless of the scheduling mechanism implemented at the available RANs. □

We remark that the algorithm naturally converges when $\widetilde{T}_{ij}(t+1) = \widetilde{T}_{ij}(t)$, i.e., when the PDAs are not willing to give away any fraction of their resource share on the RANs. However, due to network dynamics, the available resource shares on the RANs as well as the PDAs traffic demand may vary: this may trigger the PDAs to run the ANSC algorithm again and update their resource allocation.

5.6 Performance Evaluation

This section presents first the investigated network scenario that we used to derive our numerical results. Then, it evaluates the system performance and show the convergence behavior of the proposed scheme. In this context, we also compare the performance of the proposed ANSC scheme against two baseline algorithms: the Ranked Network Selection (RNS) algorithm, which implements the same idea as that proposed for network selection in [127], and the Autonomous Access Network Selection (AANS) algorithm presented in [156].

5.6.1 Simulation Environment

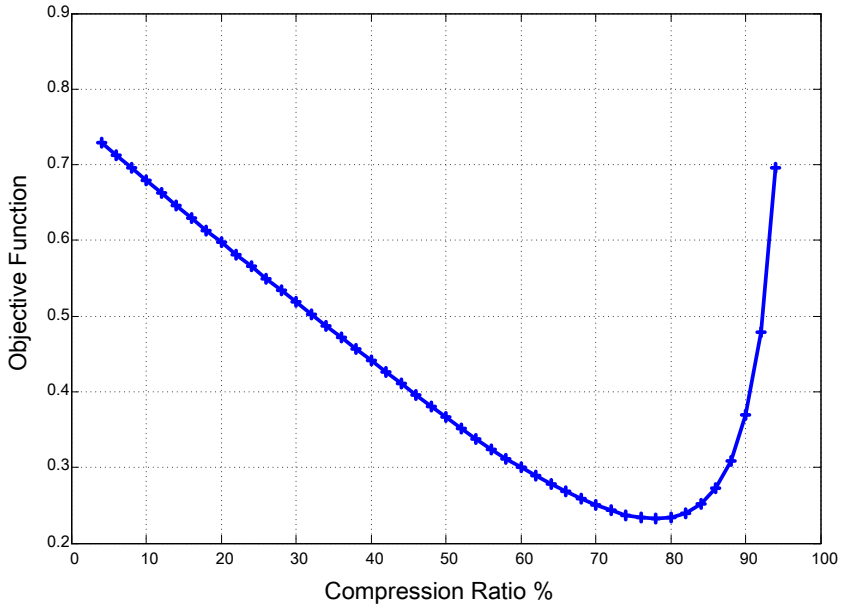
For concreteness, we consider a practical scenario of an s-health application: we consider a wireless brain monitoring system, where the PDA (i.e., smartphone) gathers EEG data from the patient using EEG Headset, then it forwards the collected data to the MHC through multi-RAN network. In this context, we exploit the EEG dataset in [51], where a PDA can capture 4096 samples of epileptic EEG data. Each raw sample is represented using 12 bits. We then consider the network topology shown in Figure 5.1, where each PDA can connect to four RANs with different characteristics. Specifically, RAN_1 with a monetary cost per bit $\varepsilon_1 = 610^{-6}$ Euro/bit and data rate $r_1 = 4$ Mbps; RAN_2 with $\varepsilon_2 = 310^{-6}$ Euro/bit and $r_2 = 2.5$ Mbps, RAN_3 with $\varepsilon_3 = 0$ Euro/bit, $r_3 = 1.5$ Mbps; RAN_4 with $\varepsilon_4 = 110^{-6}$ Euro/bit and $r_4 = 2$ Mbps. Moreover, to emphasize the tradeoff between distortion, energy consumption, latency and cost, it is assumed that $\xi_j = 0, \forall j \in M$. As far as the channel dynamics are concerned, flat Rayleigh fading is assumed, with Doppler frequency of 0.1 Hz. The other physical layer parameters over the available RANs are set to: noise spectrum density $N_0 = -174$ dBm, bandwidth $W = 0.5$ MHz, and path loss attenuation $\sigma = 3.6 * 10^{-6}$.

5.6.2 Simulation Results

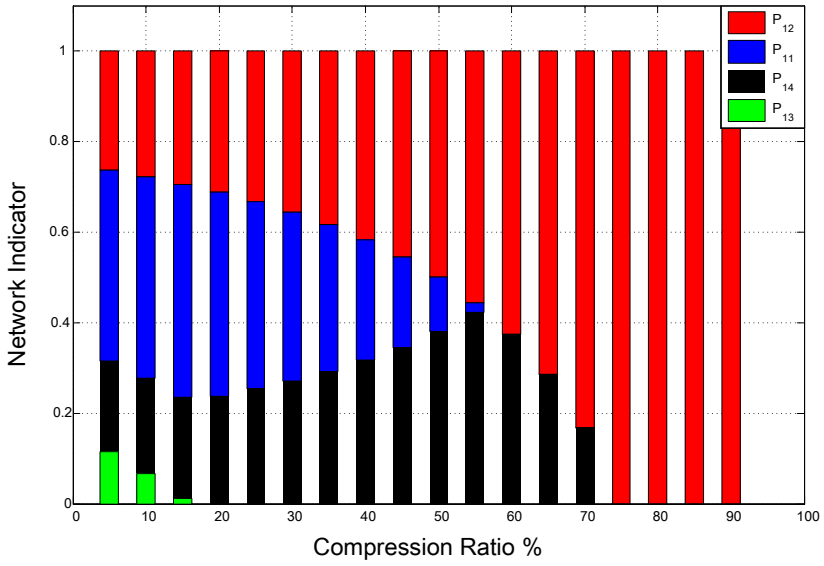
First, in order to assess the importance of optimizing both network selection and the compression ratio, Figure 5.3 depicts the value of the objective function in (5.4) as the compression ratio κ_i varies, when $\alpha_i = \beta_i = \gamma_i = \delta_i = 0.25$. One can clearly see that with increasing κ_i, the length of the transmitted data decreases; hence, initially the value of the objective function decreases as well. However, beyond a certain value, distortion becomes dominant and the value of the objective function starts increasing. Using a high compression ratio enables PDAs to decrease their load on costly networks and stick to low-cost networks, as shown by Figure 5.3-(b). This further confirms that, in order to optimize performance, it is important to jointly consider network selection and adaptive compression.

Next, we compare the performance of the proposed ANSC scheme against the two baseline algorithms. The RNS algorithm computes a score for each of the candidate RANs, using the same utility function as U_{ij}. It then selects the network with the lowest score as a target network [127]. For the sake of fairness, here we enhance RNS with the adaptive compression optimization in (5.13) to obtain the optimal value of compression ratio κ_i. In the AANS algorithm, instead, we fix κ_i to a certain value, and determine the optimal RAN(s) by solving the optimization problem in (5.9) [156]. Furthermore, we assess the ability of the tested schemes to adapt to network dynamics, in particular, we assume that the number of PDAs that can access the available RANs varies over time, as shown in Figure 5.4-(b). As expected, Figure 5.4-(a) shows that, when the number of PDAs decreases, the resource share for a generic PDA grows and the value of the aggregate objective functions drops for all schemes. The opposite occurs when new PDAs join the network. Interestingly, the network quickly adapts to any change in the scenario by assigning more or less resources to the PDAs and swiftly reaching convergence to the optimum. In all cases, however, ANSC provides the best performance.

Figure 5.5 presents the value of different performance metrics when ANSC, AANS and RNS are adopted, and the values of the corresponding network indicators P_{ij}'s over the different RANs. Here, we fixed the resource share available to each PDA to be $T_{ij} = T_j/N_j \ \forall j$. We remark that RNS selects only one network (the one with the lowest score), thus in this case P_{ij} takes a value equal to either 0 or 1 (see Figure 5.5-(H)). Our scheme and AANS instead take different candidate networks into account and select the optimal RAN(s) that minimize the PDA's

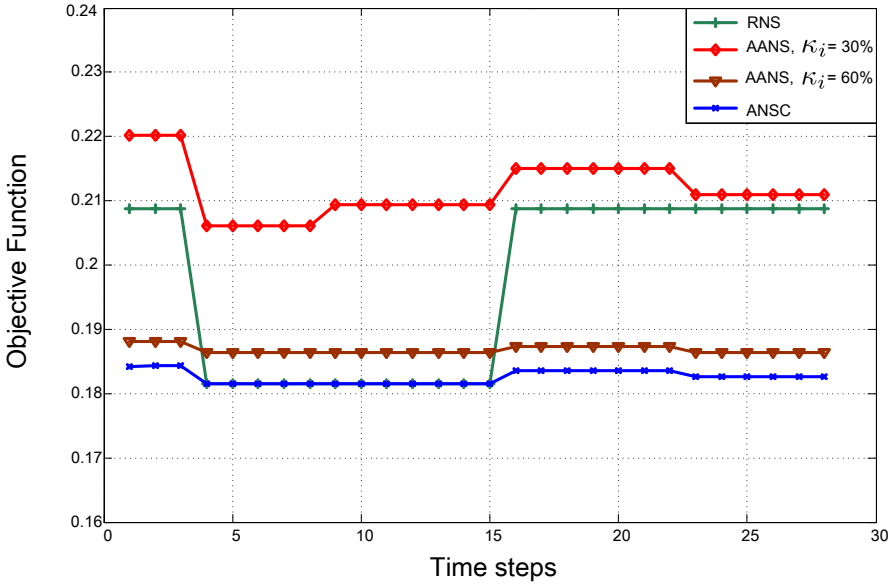


(a)

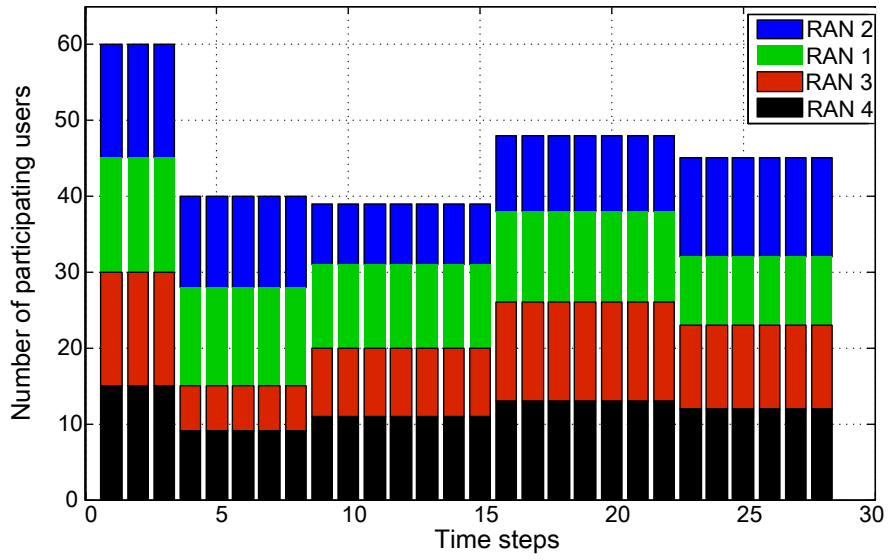


(b)

Fig. 5.3 Value of the objective function (a), and of P_{ij} (network indicators), as the compression ratio, hence distortion, varies.



(a)



(b)

Fig. 5.4 Temporal evolution of the system performance: aggregate objective function (a) as the number of PDAs varies.

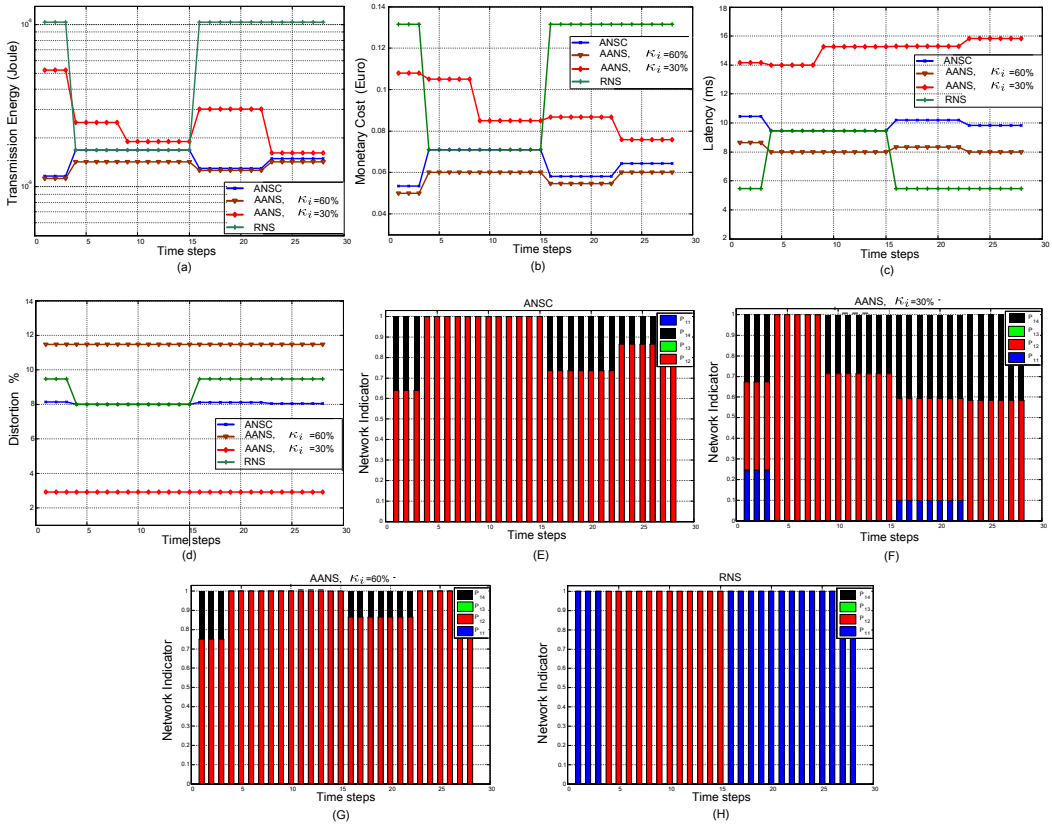


Fig. 5.5 A comparison of different performance metrics and network indicators under ANSC, AANS and RNS, with varying number of PDAs/users at each RAN.

aggregate objective, i.e., P_{ij} can take any value between 0 and 1. It follows that PDAs can transmit using different RANs simultaneously instead of being limited to one RAN only (see Figure 5.5-(E)). This enables PDAs to decrease their load on costly networks and distribute it on low-cost networks, which results in a reduced energy consumption and monetary cost (see Figure 5.5-(a),(b)).

Importantly, unlike AANS that uses a fixed compression ratio, our scheme finds the compression ratio corresponding to the optimal tradeoff between different performance metrics. Herein, we consider two possible values of compression ratio for AANS: a low κ_i, namely 30%, and a high κ_i, namely 60%. The former results to be lower than the optimal compression ratio obtained with ANSC, thus it leads to more transmitted bits. As a consequence, AANS gives a higher energy consumption and monetary cost (see Figure 5.5-(a),(b)), as it increases the amount of data transmitted on costly networks (see Figure 5.5-(F)). On the contrary, the latter value ($\kappa_i = 60\%$) is higher than the optimum. Despite the decrease in transmission

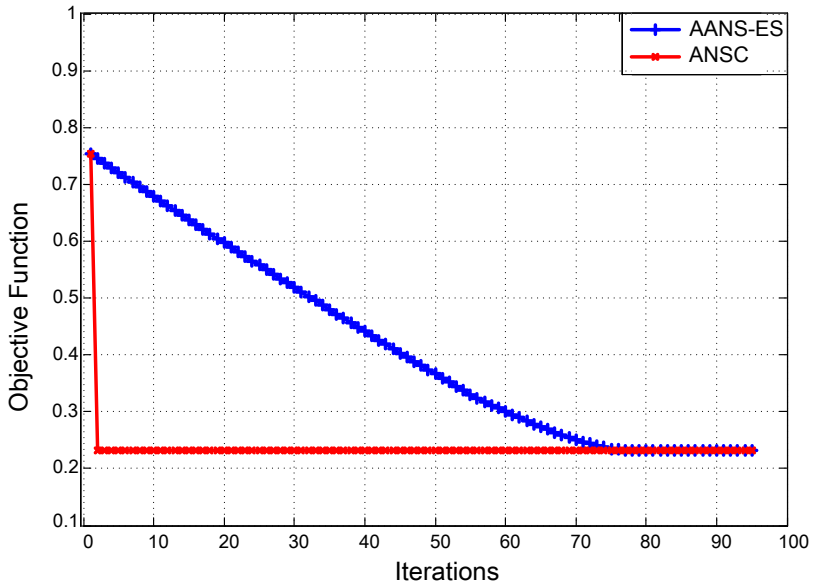


Fig. 5.6 Convergence behavior of the proposed ANSC scheme and of AANS with exhaustive search.

energy, monetary cost, and latency due to the smaller amount of transmitted data (see Figure 5.5-(a),(b),(C)), AANS leads to a higher objective function because of the large distortion (see Figure 5.5-(d)). Thus, from Figure 5.4 and Figure 5.5, we can conclude that our ANSC scheme leads to the optimal tradeoff among the target performance metrics, while other presented algorithms focus on one or more performance metric at the expense of the others.

Finally, Figure 5.6 depicts the convergence behavior of the ANSC scheme, compared to AANS. In this case, we combine AANS with exhaustive search (AANS-ES) so as to iteratively solve the optimization problem in (5.4) and find the optimal P_{ij}'s for each κ_i. Specifically, in AANS-ES, initially $\kappa_i = 0$, then it is incremented by a small quantity at every iteration. On the contrary, recall that ANSC leverages the problem decomposition into two sub-optimization problems. Although the mechanism exploited by ANSC is iterative, we observe that only three iterations are needed in order to reach convergence, compared to AANS-ES.

5.7 Summary

In this chapter, we considered a heterogeneous s-health system, where multiple radio access technologies can be simultaneously used by PDAs, in order to enhance the performance of data transfer over the wireless medium. We proposed a dynamic networks association mechanism that enables energy-efficient and high quality patient health monitoring by targeting jointly RANs selection and data compression. In the proposed scheme, the energy consumption, application QoS requirements, and monetary cost are considered as main performance metrics and integrated into a multi-objective optimization problem. We proved that the optimal solution to the problem can be obtained analytically by decomposing the problem into two sub-problems. The two sub-problems have low complexity and allow for a swift solution at the PDAs (only three iterations are needed in order to reach convergence). Our simulation results depict the efficiency of the proposed scheme and its ability to grasp the conflicting nature of users' objectives while achieving an excellent balance between them.

Chapter 6

Conclusion and Future Work

In this thesis, we presented our vision of an s-health system leveraging the multi-access edge computing paradigm. Such an approach can indeed boost the system performance by efficiently handling the enormous amount of data generated by sensors as well as medical devices at the edge of the network, and addressing the limited energy capabilities of such devices. Edge-based processing like compression and event detection can greatly reduce the amount of data transferred toward the cloud, thus removing one of the major bottlenecks in s-health systems. In this context, we presented effective approaches that can be implemented at the edge, so as to ensure short response time, efficient processing, and minimal energy and bandwidth consumption. These approaches include:

- Proposing a MEC-based architecture for satisfying s-health requirements leveraging the benefits of edge computing and context-aware optimization approaches.
- Introducing a smart health monitoring system for detecting a patient's state that exploits feature extraction and fuzzy classification to provide high accuracy, while being suitable for implementation using mobile user devices. Then, depending on the patient's state, the proposed system can exploit different data reduction techniques, in order to reduce the amount of transmitted data. In particular, under normal patient's conditions, a significant amount of energy can be saved by transmitting properly compressed data, or by sending only the most representative EEG features that are pertinent to seizures detection. Our experimental results show that, the proposed s-health system has proven

its scalability and efficiency in handling large volumes of acquired data, extending battery lifetime by 60%, and decreasing average transmission delay by 90% with respect to conventional remote monitoring systems that ignore data processing at the edge.

- Proposing an efficient EEG-based transceiver design that decreases amount of transmitted data taking into consideration the characteristics of the acquired data, while maintaining application's QoS requirements. The advantages of the proposed transceiver design are: (i) it's compatibility with the current and future standards, by leveraging well-known OFDM transceiver architecture with slight modifications and without adding significant overhead, (ii) high compression efficiency, with the remarkable result of 50% compression ratio at zero distortion and sample error rate.

- Proposing an Energy-Cost-Distortion solution, which leverages the advantages of in-network processing and medical data adaptation to optimize the transmission energy consumption and the cost of using network services. Furthermore, we present a distributed multi-objectives solution, which is suitable for heterogeneous s-health systems with variable network size. Our solution leverages Lagrangian duality theory to obtain the best tradeoff among energy consumption, network cost, and signals distortion, for delay sensitive transmission of medical data. Simulation results demonstrated the efficiency of our decentralized solution, as well as its ability to adapt to varying network conditions.

- Adopting a user-centric strategy that enables each user to independently select one or more RANs to use simultaneously. The selection depends on the user's objectives (i.e., energy saving, monetary cost, service latency, and encoding distortion), and the characteristics of the available RANs (i.e., throughput, channel quality, and data rate). Indeed, we address multi-RAN selection by formulating a multi-objective optimization problem that accounts for (i) QoS (in particular, data latency and encoding distortion) requirements, (ii) monetary cost, and (ii) energy consumption. Then, the problem is analytically solved using optimization decomposition. Finally, the convergence behavior of the proposed algorithm to the optimal solution is analyzed and proved analytically.

In short, we argue that utilization of the spectrum across diverse radio technologies is highly recommended for enhancing network capacity and quality of service for emerging applications such as remote health monitoring. Furthermore, it is also recommended to integrate wireless network components, characteristics of the acquired data, and high level requirements of the considered application, in order to provide sustainable, energy-efficient and high-quality services for smart health systems.

In what follows, we discuss three main directions that represent interesting lines for future research.

1. **Privacy and security**: Great potential of s-health system can only be achieved if individuals are confident about the privacy of their health-related information, and providers are confident about the security of gathered data. However, ensuring privacy and security is not straightforward. Wireless medical devices are typically susceptible to various types of threats, such as patient tracking and relaying, as well as denial of service attacks, which violate confidentiality and integrity of the devices. Data processing algorithms and data storage may also be subject to attacks. Below, we discuss some challenges and opportunities that MEC poses in this respect.

   First is the ownership of the collected data from the patients. Storing the data at the patients' proximity, where it is collected, and enabling the patients to fully own the data is a better solution for privacy protection. Also, the patient will be able to control if the data should be stored at the edge or transmitted to the cloud after removing or hiding some of the private information from the data.

   Second is the tradeoff between increasing security level and QoS. Increased security through strong cryptographic algorithms or effective key management schemes [157][158], adds more processing and additional overhead at the edge, which may have a significantly adverse impact on QoS, especially for real-time applications with strict delay and throughput requirements. This imposes an essential need to design joint QoS and security mechanisms for s-health applications that maximize QoS, while meeting applications' security requirements.

2. **Collaborative edge**: Healthcare requires data sharing and collaboration among different stakeholders in multiple domains. However, sharing of data owned by a stakeholder rarely happens due to privacy concerns and the high cost of data transfer. In this context, collaborative edge, which connects the edges of multiple stakeholders that are geographically distributed (such as hospitals, centers for disease control and prevention, pharmacies, and insurance companies), is beneficial in threefold. First, it provides distributed data sharing among different stakeholders at low cost, thanks to computation and processing at the participant edges. Second, in the case of remote monitoring, it enables patients to forward their medical data to the cloud through other users/edge nodes. This also improves spectrum and energy efficiency and allows data transferring even in geographically remote areas by exploiting Device-to-Device (D2D) data transfer [159][160]. Third, it enables a patient's edge node to directly connect to the nearest hospital's edge in the proximity for continuous monitoring, without the need of going through the cloud. This helps to increase monitoring efficiency, reduce energy consumption and operational cost, as well as provide high-quality services.

3. **Combining heterogeneous sources of information**: Various sources of information are used in s-health systems for efficient monitoring, hence, leveraging advanced multimodal data processing techniques for combining these sources of information at the edge is a promising trend toward automating supervision and remote monitoring tasks. Multi-radio and multi-technology Edge gateway can be considered to locally process data coming from various data sources. However, several challenges remain open when it comes to the s-health systems with hybrid sensing sources. First, in terms of multiple modalities, it is not straightforward to incorporate and transmit multiple data streams in s-health systems, where power consumption is a limiting factor; indeed, transmission of highly informative biosignals (e.g., EEG, EMG, and electrocardiogram) is an energy hungry process for battery-operated devices. Second, signals artifacts arise from internal sources, e.g., muscle activities and movements, as well as from external sources related to noise, interference, and signals offset, which have critical implications on data quality [161].

   In this context, adopting a MEC-based s-health system architecture would be beneficial in two ways. First, it permits to address system complexity associated with such heterogeneous and variable data-stream inputs. This is

done through implementing multimodal in-network processing techniques that yield the correlation between different modalities, in addition to the temporal correlation within each modality [162]. Moreover, a MEC-based architecture enables extracting high level application-based features at the edge rather than the cloud. By doing so, a MEN can send a limited number of the extracted features, or the obtained correlations, instead of transmitting either the original or the compressed data. Second, advanced signal processing for artifact removal can be incorporated at the edge, in order to improve signals quality before transmission.

References

[1] M. Kim and S. Chang. A consumer transceiver for long-range IoT communications in emergency environments. *IEEE Transactions on Consumer Electronics*, 62(3):226–234, August 2016.

[2] World Health Organization. mHealth: New horizons for health through mobile technologies. Global Observatory for eHealth Series, 2011.

[3] S. Latif, R. Rana, J. Qadir, A. Ali, M. A. Imran, and M. S. Younis. Mobile health in the developing world: Review of literature and lessons from a case study. *IEEE Access*, 5:11540–11556, 2017.

[4] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. A. Perez-martinez, R. D. Pietro, D. N. Perrea, and A. Martinez-Balleste. Smart health: A context-aware health paradigm within smart cities. *IEEE Communications Magazine*, 52(8):74–81, Aug 2014.

[5] Christoph Thuemmler and Chunxue Bai. Health 4.0: application of industry 4.0 design principles in future Asthma management. *Springer International Publishing*, pages 23–37, 2017.

[6] Cisco global cloud index: Forecast and methodology, 2014-2019 white paper. 2014.

[7] Ayush Bansal, Sunil Kumar, Anurag Bajpai, Vijay N. Tiwari, Mithun Nayak, Shankar Venkatesan, and Rangavittal Narayanan. Remote health monitoring system for detecting cardiac disorders. *IET Systems Biology*, 9(6):309–314, Dec 2015.

[8] Priyanka Kakria, N. K. Tripathi, and Peerapong Kitipawang. A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors. *International Journal of Telemedicine and Applications*, 2015, 2015.

[9] S. Ansari, N. Farzaneh, M. Duda, K. Horan, H. B. Andersson, Z. D. Goldberger, B. K. Nallamothu, and K. Najarian. A review of automated methods for detection of myocardial ischemia and infarction using electrocardiogram and electronic health records. *IEEE Reviews in Biomedical Engineering*, 10:264–298, 2017.

[10] M. Alhussein. Monitoring Parkinson's disease in smart cities. *IEEE Access*, 5:19835–19841, 2017.

[11] M.A. Hassan, A.S. Malik, D. Fofi, N. Saad, B. Karasfi, Y.S. Ali, and F. Meriaudeau. Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control*, 38:346–360, 2017.

[12] Aatif M. Husain and Saurabh R. Sinha. *Continuous EEG Monitoring*. Springer International Publishing, first edition, 2017.

[13] A. H. Gee, R. Barbieri, D. Paydarfar, and P. Indic. Predicting Bradycardia in preterm infants using point process analysis of heart rate. *IEEE Transactions on Biomedical Engineering*, 64(9):2300–2308, Sept 2017.

[14] Lasitha S. Vidyaratne and Khan M. Iftekharuddin. Real-time epileptic seizure detection using EEG. *IEEE Transactions on Neural systems and Rehabilitation Engineering*, 25(11), Nov. 2017.

[15] F. A. Khan, N. A. H. Haldar, A. Ali, M. Iftikhar, T. A. Zia, and A. Y. Zomaya. A continuous change detection mechanism to identify anomalies in ecg signals for wban-based healthcare environments. *IEEE Access*, 5:13531–13544, 2017.

[16] Ali Hakan Isik, Inan Guler, and Melahat Uzel Sener. A low-cost mobile adaptive tracking system for chronic pulmonary patients in home environment. *Telemed J E Health*, 19(1), Jan 2017.

[17] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5):637–646, Oct 2016.

[18] Multi-access edge computing. *http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing*, Accessed on July 2017.

[19] Tarik Taleb, Konstantinos Samdanisz, Badr Mada, Hannu Flinck, Sunny Dutta, and Dario Sabella. On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture & orchestration. *IEEE Communications Surveys & Tutorials*, 2017.

[20] R.F. Yazicioglu, T. Torfs, P. Merken, J. Penders, V. Leonov, R. Puers, B. Gyselinckx, and C. van Hoof. Ultra-low-power biopotential interfaces and their applications in wearable and implantable systems. *Microelectron. J.*, pages 1313–1321, 2009.

[21] Alaa Awad Abdellatif, Ahmed Emam, C. Chiasserini, A. Mohamed, Ali Jaoua, and Rabab Ward. Edge-based compression and classification for smart healthcare systems: Concept, implementation and evaluation. *Expert Systems with Applications*, 2018.

[22] A. A. Abdellatif, A. Mohamed, and C. Chiasserini. Automated class-based compression for real-time epileptic seizure detection. In *Wireless Telecommunications Symposium (WTS)*, pages 1–6, April 2018.

[23] Thurman et al. Standards for epidemiologic studies and surveillance of epilepsy. *Epilepsia*, 52(7):2–26, 2011.

[24] Hojjat Adeli, Samanwoy Ghosh-Dastidar, and Nahid Dadmehr. A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy. *IEEE Transactions on Biomedical Engineering, vol. 54, NO. 2*, February 2007.

[25] M.R. Yuce, S.W.P. Ng, N.L. Myo, J.Y. Khan, and W. Liu. Wireless body sensor network using medical implant band. *J. Medical Systems*, 31(6):467–474, 2007.

[26] F. A. Kraemer, A. E. Braten, N. Tamkittikhun, and D. Palma. Fog computing in healthcare–a review and discussion. *IEEE Access*, 5:9206–9222, 2017.

[27] B. Xu, L. D. Xu, H. Cai, C. Xie, J. Hu, and F. Bu. Ubiquitous data accessing method in IoT-Based information system for emergency medical services. *IEEE Transactions on Industrial Informatics*, 10(2):1578–1586, May 2014.

[28] Z. Sheng, C. Mahapatra, C. Zhu, and V. C. M. Leung. Recent advances in industrial wireless sensor networks toward efficient management in IoT. *IEEE Access*, 3:622–637, 2015.

[29] S. Sarkar and S. Misra. Theoretical modelling of fog computing: a green computing paradigm to support IoT applications. *IET Networks*, 5(2):23–29, 2016.

[30] Mohamed EL Menshawy, Abdelghani Benharref, and Mohamed Serhani. An automatic mobile-health based approach for eeg epileptic seizures detection. *Expert Systems with Applications*, 42:7157–7174, 2015.

[31] A. V. Dastjerdi and R. Buyya. Fog computing: Helping the internet of things realize its potential. *Computer*, 49(8):112–116, Aug 2016.

[32] P. Pace, G. Aloi, R. Gravina, G. Caliciuri, G. Fortino, and A. Liotta. An edge-based architecture to support efficient applications for healthcare industry 4.0. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2018.

[33] L. Cerina, S. Notargiacomo, M. G. Paccanit, and M. D. Santambrogio. A fog-computing architecture for preventive healthcare and assisted living in smart ambients. In *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*, pages 1–6, 2017.

[34] L. A. Rossi, B. Krishnamachari, and C. C. J. Kuo. Energy efficient data collection via supervised in-network classification of sensor data. In *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 33–42, May 2016.

[35] M. A. Alsheikh, S. Lin, D. Niyato, and H. P. Tan. Machine learning in wireless sensor networks: Algorithms, strategies, and applications. *IEEE Communications Surveys Tutorials*, 16(4):1996–2018, Fourthquarter 2014.

[36] Bartolomeo Cosenza. Off-line control of the postprandial glycemia in type 1 diabetes patients by a fuzzy logic decision support. *Expert Systems with Applications*, 39(12):10693–10699, 2012.

[37] F. Tatari, M. Akbarzadeh, and A. Sabahi. Fuzzy probabilistic multi-agent system for breast cancer risk assessment and insurance premium assignment. *Journal of Biomedical Informatics*, 45(6):1021–1034, December 2012.

[38] Jumanah A Al-Dmour, Assim Sagahyroon, AR Al-Ali, and Salah Abusnana. A fuzzy logic–based warning system for patients classification. *Health Informatics Journal*, pages 1–21, Nov. 2017.

[39] M.J.P. Castanho, F. Hernandes, A.M. De Ré, S. Rautenberg, and A. Billis. Fuzzy expert system for predicting pathological stage of prostate cancer. *Expert Systems with Applications*, 40(2):466–470, February 2013.

[40] M. Hooshmand, D. Zordan, D. Del Testa, E. Grisan, and M. Rossi. Boosting the battery life of wearables for health monitoring through the compression of biosignals. *IEEE Internet of Things Journal*, 4(5):1647–1662, Oct 2017.

[41] G. Antoniol and P. Tonella. EEG data compression techniques. *IEEE Trans. Biomed. Eng.*, 4:105–114, 1997.

[42] A. A. Funmilola, D. F. Olusayo, and A. A. Michael. Comparative analysis between discrete cosine transform and wavelet transform techniques for medical image compression. In *International Conference on Computer Vision and Image Analysis Applications (ICCVIA)*, pages 1–6, Jan 2015.

[43] I. Capurro, F. Lecumberry, Á. Martín, I. Ramírez, E. Rovira, and G. Seroussi. Efficient sequential compression of multichannel biomedical signals. *IEEE Journal of Biomedical and Health Informatics*, 21(4):904–916, July 2017.

[44] Joyce Chiang and Rabab K. Ward. Energy-efficient data reduction techniques for wireless seizure detection systems. *Sensors, 14(2)*, pages 2036–2051, 2014.

[45] Yatindra Kumar, M.L. Dewal, and R.S. Anand. Epileptic seizure detection using dwt based fuzzy approximate entropy and support vector machine. *Original Research Article Neurocomputing*, 133:271–279, 2014.

[46] Lan-Lan Chen, Jian Zhang, Jun-Zhong Zou, Chen-Jie Zhao, and Gui-Song Wang. A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection. *Biomedical Signal Processing and Control*, 10:1–10, 2014.

[47] Epilepsy detector application. *http://www.epdetect.com/*, 2017.

[48] A. Helmy and A. Helmy. Seizario: Novel mobile algorithms for seizure and fall detection. In *IEEE Globecom Workshops (GC Wkshps)*, pages 1–6, Dec 2015.

[49] Berg AT, Shinnar S, Levy SR, Testa FM, Smith-Rapaport S, Beckerman B, and Ebrahimi N. Two-year remission and subsequent relapse in children with newly diagnosed epilepsy. *Epilepsia*, 42(12):1553–1562, 2001.

[50] Englander and Jeffrey et al. Seizures after traumatic brain injury. *Archives of physical medicine and rehabilitation*, 95(6):1223–1224, 2014.

[51] RG Andrzejak, K Lehnertz, C Rieke, F Mormann, P David, and CE Elger. Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Phys. Rev. E, 64, 061907, (2001)*, 2001.

[52] Dennis Tkach, He Huang, and Todd Kuiken. Study of stability of time-domain features for electromyographic pattern recognition. In *Journal of NeuroEngineering and Rehabilitation, vol. 7*, May 2010.

[53] R. Hussein, A. Mohamed, K. Shahan, and A. A. Mohamed. Eeg feature extraction and selection techniques for epileptic detection: A comparative study. In *IEEE Symposium on Computers Informatics (ISCI)*, pages 170–175, April 2013.

[54] Angkoon Phinyomark, Chusak Limsakul, and Pornchai Phukpattaranont. A novel feature extraction for robust EMG pattern recognition. *Journal of Computing*, 1:71–80, 2009.

[55] John G. Proakis and Dimitris G. Manolakis. *"Digital signal processing: principles, algorithms and applications*. Pearson Prentice Hall, fourth edition, 2007.

[56] William L. Maner, Robert E. Garfield, Holger Maul, Gayle Olson, and George Saade. Predicting term and preterm delivery with transabdominal uterine electromyography. *Obstetrics and Gynecology*, 101(6):1254–1260, 2003.

[57] Mitul Kumar Ahirwal and Narendra D londhe. Power spectrum analysis of EEG signals for estimating visual attention. *International Journal of Computer Applications*, 42(15):34–40, March 2012.

[58] R. Hussein, A. Mohamed, and M. Alghoniemy. Energy-efficient on-board processing technique for wireless epileptic seizure detection systems. In *International Conference on Computing, Networking and Communications (ICNC)*, pages 1116–1121, Feb 2015.

[59] M. Maddouri, S. Elloumi, and A. Jaoua. An incremental learning system for imprecise and uncertain knowledge discovery. *Journal of Information Sciences 109*, pages 149–164, 1998.

[60] Jihad M. Alja'am, Ali Jaoua, Ahmad Hasnah, F. Hassan, H. Mohamed, T. Mosaid, H. Saleh, F. Abdullah, and H. Cherif. Text summarization based on conceptual data classification. *International Journal of Information Technology and Web Engineering*, pages 22–36, 2006.

[61] Samir Elloumi, Jihad Mohamad Jaam, Ahmad Hasnah, Ali Jaoua, and Ibtissem Nafkha. A multi-level conceptual data reduction approach based on the lukasiewicz implication. *Information Sciences 163*, pages 253–262, 2004.

[62] Didier Dubois and Henri Prade. *Fundamentals of Fuzzy Sets*. Springer Science & Business Media, first edition, 2000.

[63] V. Novak. *Fuzzy Sets and their Applications*. Adam Hilger, Bristol, 1989.

[64] Muhammad Sarfraz. *Computer-Aided Intelligent Recognition Techniques and Applications*. John Wiley & Sons, second edition, 2005.

[65] Raoudha Khcherif, Mohamed Mohsen Gammoudi, and Ali Jaoua. Using difunctional relations in information organization. *Information Sciences 125*, pages 153–166, 2000.

[66] Alaa Awad, Amr Mohamed, Amr A. El-Sherif, and Omar A. Nasr. Interference-aware energy-efficient cross-layer design for healthcare monitoring applications. *Comput. Netw., vol. 74*, pages 64–77, December 2014.

[67] Ramy Hussein, Amr Mohamed, and Masoud Alghoniemy. Scalable real-time energy-efficient EEG compression scheme for wireless body area sensor network. *Biomedical Signal Processing and Control 19*, pages 122–129, 2015.

[68] Stephane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, third edition, 2008.

[69] IEEE. "wireless LAN medium access control (MAC) and physical layer (PHY) specifications. *IEEE Standard 802.11*, 2007.

[70] Svetlana S. Aksenova. Weka explorer tutorial. *School of Engineering and Computer Science California State University*, 2004.

[71] Analyzing power use with battery historian. In *https://developer.android.com/topic/ performance/power/battery-historian.html, last visited*, April 2018.

[72] A. Awad, A. Mohamed, and A. A. El-Sherif. Energy efficient cross-layer design for wireless body area monitoring networks in healthcare applications. In *IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pages 1484–1489, Sept 2013.

[73] A. A. Abdellatif, M. G. Khafagy, A. Mohamed, and C. Chiasserini. Eeg-based transceiver design with data decomposition for healthcare iot applications. *IEEE Internet of Things Journal*, pages 1–1, 2018.

[74] Y. Chae, J. Jeong, and S. Jo. Toward brain-actuated humanoid robots: Asynchronous direct control using an eeg-based bci. *IEEE Transactions on Robotics*, 28(5):1131–1144, Oct 2012.

[75] F. Duan, D. Lin, W. Li, and Z. Zhang. Design of a multimodal eeg-based hybrid bci system with visual servo module. *IEEE Transactions on Autonomous Mental Development*, 7(4):332–341, Dec 2015.

[76] A. J. C. Trappey, C. V. Trappey, U. H. Govindarajan, J. J. Sun, and A. C. Chuang. A review of technology standards and patent portfolios for enabling cyber-physical systems in advanced manufacturing. *IEEE Access*, 4:7356–7382, 2016.

[77] T. Moy, L. Huang, W. Rieutort-Louis, C. Wu, P. Cuff, S. Wagner, J. C. Sturm, and N. Verma. An eeg acquisition and biomarker-extraction system using low-noise-amplifier and compressive-sensing circuits based on flexible, thin-film electronics. *IEEE Journal of Solid-State Circuits*, 52(1):309–321, Jan 2017.

[78] D. Craven, B. McGinley, L. Kilmartin, M. Glavin, and E. Jones. Compressed sensing for bioelectric signals: A review. *IEEE Journal of Biomedical and Health Informatics*, 19(2):529–540, March 2015.

[79] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, Jan 2017.

[80] Thomas Strohmer. Measure what should be measured: Progress and challenges in compressive sensing. *IEEE Signal Processing Letters*, 19(12):887–893, 2012.

[81] S. Acciarito, G.C. Cardarilli, L. Di Nunzio, R. Fazzolari, G.M. Khanal, and M. Re. Compressive sensing reconstruction for complex system: A hardware/software approach. *Applications in Electronics Pervading Industry, Environment and Society, Springer, Cham*, 429:192–200, 2016.

[82] B. L. Sturm and M. G. Christensen. Comparison of orthogonal matching pursuit implementations. *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 220–224, Aug 2012.

[83] C. Hegde, A. C. Sankaranarayanan, W. Yin, and R. G. Baraniuk. Numax: A convex approach for learning near-isometric linear embeddings. *IEEE Transactions on Signal Processing*, 63(22):6109–6121, Nov 2015.

[84] Y. Wang, X. Li, K. Xu, F. Ren, and H. Yu. Data-driven sampling matrix boolean optimization for energy-efficient biomedical signal acquisition by compressive sensing. *IEEE Transactions on Biomedical Circuits and Systems*, 11(2):255–266, April 2017.

[85] M. Hooshmand, D. Zordan, D. Del Testa, E. Grisan, and M. Rossi. Boosting the battery life of wearables for health monitoring through the compression of biosignals. *IEEE Internet of Things Journal*, 4(5):1647–1662, Oct 2017.

[86] D. Del Testa and M. Rossi. Lightweight lossy compression of biometric patterns via denoising autoencoders. *IEEE Signal Processing Letters*, 22(12):2304–2308, Dec 2015.

[87] N. Sriraam and C. Eswaran. Performance evaluation of neural network and linear predictors for near-lossless compression of EEG signals. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):87–93, Jan 2008.

[88] S. Lee, J. Kim, and M. Lee. A real-time ecg data compression and transmission algorithm for an e-health device. *IEEE Transactions on Biomedical Engineering*, 58(9):2448–2455, Sept 2011.

[89] H. Daou and F. Labeau. Dynamic dictionary for combined EEG compression and seizure detection. *IEEE Journal of Biomedical and Health Informatics*, 18(1):247–256, Jan 2014.

[90] G. Xu, J. Han, Y. Zou, and X. Zeng. A 1.5-D multi-channel EEG compression algorithm based on NLSPIHT. *IEEE Signal Processing Letters*, 22(8):1118–1122, Aug 2015.

[91] E. S. G. Carotti, J. C. De Martin, R. Merletti, and D. Farina*. Compression of multidimensional biomedical signals with spatial and temporal codebook-excited linear prediction. *IEEE Transactions on Biomedical Engineering*, 56(11):2604–2610, Nov 2009.

[92] S. Brandstätter and M. Huemer. A novel MPSoC interface and control architecture for multistandard rf transceivers. *IEEE Access*, 2:771–787, 2014.

[93] E.O. Schweiter and T.M. Minteer. Fiber-optic transceiver for combined serial data and time code communication, May 18 1999. US Patent 5,905,758.

[94] G.D. Knox. Wireless music and data transceiver system, June 29 2004. US Patent 6,757,913.

[95] P. Guan, D. Wu, T. Tian, J. Zhou, X. Zhang, L. Gu, A. Benjebbour, M. Iwabuchi, and Y. Kishiyama. 5g field trials: Ofdm-based waveforms and mixed numerologies. *IEEE Journal on Selected Areas in Communications*, 35(6):1234–1243, June 2017.

[96] D. C. Alves, G. S. da Silva, E. R. de Lima, C. G. Chaves, D. Urdaneta, T. Perez, and M. Garcia. Architecture design and implementation of key components of an OFDM transceiver for IEEE 802.15.4g. *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 550–553, May 2016.

[97] RG Andrzejak, K Lehnertz, C Rieke, F Mormann, P David, and CE Elger. Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Phys. Rev. E, 64, 061907, (2001)*, 2001.

[98] Tzi-Dar Chiueh and Pei-Yun Tsai. *OFDM Baseband Receiver Design for Wireless Communications*. John Wiley & Sons, first edition, 2007.

[99] Yun Li, Yunhao Yuan, Xin Guo, Yan Sheng, and Ling Chen. A fast algorithm for generating concepts. In *International Conference on Information and Automation*, pages 1728–1733, June 2008.

[100] J. Trein, A. T. Schwarzbacher, B. Hoppe, and K. H. Noff. A hardware implementation of a run length encoding compression algorithm with parallel inputs. *IET Irish Signals and Systems Conference (ISSC 2008)*, pages 337–342, June 2008.

[101] A. Awad, M. Hamdy, A. Mohamed, and H. Alnuweiri. Real-time implementation and evaluation of an adaptive energy-aware data compression for wireless EEG monitoring systems. *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, pages 108–114, Aug 2014.

[102] Alaa Awad, Amr Mohamed, Carla-Fabiana Chiasserini, and Tarek Elfouly. Distributed in-network processing and resource optimization over mobile-health systems. *Journal of Network and Computer Applications*, 82:65–76, March 2017.

[103] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and L. Victor. Body area networks: A survey. *J. Mobile Netw. Appl., VOL. 16*, pages 171–193, 2011.

[104] O. D. Incel, L. van Hoesel, P. Jansen, and P. Havinga. MC-LMAC: A multi-channel MAC protocol for wireless sensor networks. *AdHoc Networks, vol. 9*, pages 73–94, January 2011.

[105] B. Otal, L. Alonso, and C. Verikoukis. Highly reliable energy-saving mac for wireless body sensor networks in healthcare systems. *IEEE Journal on Selected Areas in Communications, vol. 27, no. 4*, May 2009.

[106] A. Jain, D. Gunduz, S.R. Kulkarni, H.V. Poor, and S. Verdu. Energy-distortion tradeoffs in gaussian joint source-channel coding problems. *IEEE Transactions on Information Theory*, 58(5):3153–3168, May 2012.

[107] Yifeng He, Wenwu Zhu, and Ling Guan. Optimal resource allocation for pervasive health monitoring systems with body sensor networks. *IEEE Transactions on Mobile Computing*, 10(11):1558–1575, Nov 2011.

[108] J.G. Andrews, S. Buzzi, Wan Choi, S.V. Hanly, A. Lozano, A.C.K. Soong, and J.C. Zhang. What will 5g be? *IEEE Journal on Selected Areas in Communications*, 32(6):1065–1082, June 2014.

[109] R. Trestian, O. Ormond, and G.-M. Muntean. Energy-quality-cost tradeoff in a multimedia-based heterogeneous wireless network environment. *IEEE Transactions on Broadcasting*, 59(2):340–357, June 2013.

[110] Yiqun Wu, Y. Chen, Jie Tang, D.K.C. So, Zhikun Xu, Chih-Lin I, P. Ferrand, J.-M. Gorce, Chih-Hsuan Tang, Pei-Rong Li, Kai-Ten Feng, Li-Chun Wang, K. Borner, and L. Thiele. Green transmission technologies for balancing the

energy efficiency and spectrum efficiency trade-off. *IEEE Communications Magazine*, 52(11):112–120, Nov 2014.

[111] Jiucai Zhang, Dalei Wu, Song Ci, Haohong Wang, and Aggelos K. Katsaggelos. Power-aware mobile multimedia: a survey. *Journal of Communications*, 4(9), 2009.

[112] T. Ma, M. Hempel, D. Peng, and H. Sharif. A survey of energy-efficient compression and communication techniques for multimedia in resource constrained systems. *IEEE Communications Surveys Tutorials*, 15(3):963–972, Third 2013.

[113] Eda Akman Aydın, Omer Faruk Bay, and Inan Guler. Implementation of an embedded web server application for wireless control of brain computer interface based home environments. *Journal of Medical Systems*, 40(27), January 2016.

[114] Isik AH., Guler I., and Sener MU. A low-cost mobile adaptive tracking system for chronic pulmonary patients in home environment. *Telemedicine and e-Health*, 19(1):24–30, January 2013.

[115] Majeed-Ariss and Rabiya et al. Apps and adolescents: A systematic review of adolescents' use of mobile phone and tablet apps that support personal management of their chronic or long-term physical conditions. *Ed. Gunther Eysenbach. Journal of Medical Internet Research*, 17(12), Dec 2015.

[116] Jingting Wang et al. Smartphone interventions for long-term health management of chronic diseases: An integrative review. *Telemedicine and e-Health*, 20(6):570–583, June 2014.

[117] Jia Liu, Cathy H. Xia, Ness B. Shroff, and Hanif D. Sherali. Distributed cross-layer optimization in wireless networks: A second-order approach. *Proceedings in IEEE INFOCOM*, 2013.

[118] Sang-Seon Byun, Ilangko Balasingham, and Xuedong Liang. Dynamic spectrum allocation in wireless cognitive sensor networks: Improving fairness and energy efficiency. *IEEE Vehicular Technology Conference*, September 2008.

[119] Alaa Awad, Ramy Hussein, Amr Mohamed, and Amr A. El-Sherif. Energy-aware cross-layer optimization for EEG-based wireless monitoring applications. *IEEE Conference on Local Computer Networks (LCN)*, pages 356–363, Oct 2013.

[120] Alaa Awad, Amr Mohamed, Amr A. El-Sherif, and Omar A. Nasr. Interference-aware energy-efficient cross-layer design for healthcare monitoring applications. *Computer Networks, voL. 74*, pages 64–77, December 2014.

[121] Alaa Awad, Medhat Hamdy, Amr Mohamed, and Hussein Alnuweiri. Real-time implementation and evaluation of an adaptive energy-aware data compression for wireless EEG monitoring systems. *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE)*, pages 108–114, August 2014.

[122] A. Awad, A. Mohamed, and T. Elfouly. Energy-cost-distortion optimization for delay-sensitive m-health applications. *Wireless Telecommunications Symposium (WTS)*, pages 1–5, April 2015.

[123] REBECCA M. WARNER. *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA, US: Sage Publications, second edition, 2008.

[124] J. Cárdenas-Barrera, J. Lorenzo-Ginori, and E. Rodríguez-Valdivia. A wavelet-packets based algorithm for EEG signal compression. *Med Informatic and Internet in Med*, March 2004.

[125] IEEE standard for local and metropolitan area networks - part 15.6: Wireless body area networks. *IEEE Std 802.15.6-2012*, pages 1–271, Feb 2012.

[126] IEEE standard for local and metropolitan area networks-part 21: Media independent handover. *IEEE Std. 802.21-2008*, 2009.

[127] R. Trestian, O. Ormond, and G.-M. Muntean. Enhanced power-friendly access network selection strategy for multimedia delivery over heterogeneous wireless networks. *IEEE Transactions on Broadcasting, vol.60, no.1*, pages 85–101, March 2014.

[128] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. cambridge university press, first edition, 2003.

[129] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optimization and Eng., vol. 8, no. 1*, pages 67–127, Apr 2007.

[130] D. P. Bertsekas. *Nonlinear Programming*. Belmont, MA: Athena Scientific, second edition, 1999.

[131] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[132] M. A. Gennert and A. Yuille. Determining the optimal weights in multiple objective function optimization. *2nd International Conference on Computer Vision,*, pages 87–89, December 1988.

[133] S. Taleb, H. Hajj, and Z. Dawy. Entropy-based optimization to trade-off energy and accuracy for activity mobile sensing. *4th Annual International Conference on Energy Aware Computing Systems and Applications (ICEAC)*, pages 6–11, Dec 2013.

[134] B. Zhen et al. TG6 technical requirements document. *IEEE P802.15-08-0644-09-0006; https://mentor.ieee.org/802.15/.*

[135] J. Andrews, S. Singh, Qiaoyang Ye, Xingqin Lin, and H. Dhillon. An overview of load balancing in hetnets: old myths and open problems. *IEEE Wireless Communications*, 21(2):18–25, April 2014.

[136] A. A. Abdellatif, A. Mohamed, and C. Chiasserini. User-centric networks selection with adaptive data compression for smart health. *IEEE Systems Journal*, pages 1–11, 2018.

[137] J. Andrews, S. Singh, Qiaoyang Ye, Xingqin Lin, and H. Dhillon. An overview of load balancing in hetnets: old myths and open problems. *IEEE Wireless Communications*, 21(2):18–25, April 2014.

[138] J. Wu, J. Liu, Z. Huang, C. Du, H. Zhao, and Y. Bai. Intelligent network selection for data offloading in 5g multi-radio heterogeneous networks. *China Communications*, 12:132–139, December 2015.

[139] D. Niyato and E. Hossain. Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach. *IEEE Transactions on Vehicular Technology*, 58(4):2008–2017, May 2009.

[140] P. Naghavi et al. Learning RAT selection game in 5G heterogeneous networks. *IEEE Wireless Communications Letters*, 5(1):52–55, Feb 2016.

[141] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong. An mdp-based vertical handoff decision algorithm for heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 57(2):1243–1254, March 2008.

[142] M. El Helou, M. Ibrahim, S. Lahoud, K. Khawam, D. Mezher, and B. Cousin. A network-assisted approach for rat selection in heterogeneous cellular networks. *IEEE Journal on Selected Areas in Communications*, 33(6):1055–1067, June 2015.

[143] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews. User association for load balancing in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, 12(6):2706–2716, June 2013.

[144] Q. Wu, Z. Du, P. Yang, Y. D. Yao, and J. Wang. Traffiic-aware online network selection in heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 65(1):381–397, Jan 2016.

[145] K. Habak, K. A. Harras, and M. Youssef. What goes around comes around: Mobile bandwidth sharing and aggregation. In *IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pages 37–45, Oct 2015.

[146] T. Dreibholz, E. P. Rathgeb, I. Rungeler, R. Seggelmann, M. Tuxen, and R. R. Stewart. Stream control transmission protocol: Past, current, and future standardization activities. *IEEE Communications Magazine*, 49(4):82–88, April 2011.

[147] J. Wu, C. Yuen, M. Wang, and J. Chen. Content-aware concurrent multipath transfer for high-definition video streaming over heterogeneous wireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 27(3):710–723, March 2016.

[148] Gubong Lim, Cong Xiong, L.J. Cimini, and G.Y. Li. Energy-Efficient Resource Allocation for OFDMA-Based Multi-RAT Networks. *IEEE Transactions on Wireless Communications*, 13(5):2696–2705, May 2014.

[149] M. Fadel, A.S. Ibrahim, and H. Elgebaly. QoS-Aware Multi-RAT resource allocation with minimum transmit power in multiuser OFDM system. In *IEEE Globecom Workshops (GC Wkshps)*, pages 670–675, Dec 2012.

[150] M. Gerasimenko, N. Himayat, Shu-Ping Yeh, S. Talwar, S. Andreev, and Y. Koucheryavy. Characterizing performance of load-aware network selection in multi-radio (WiFi/LTE) heterogeneous networks. In *IEEE Globecom Workshops (GC Wkshps)*, pages 397–402, Dec 2013.

[151] L. Wang and G. Kuo. Mathematical modeling for network selection in heterogeneous wireless networks-A tutorial. *IEEE Communications Surveys Tutorials*, 15(1):271–292, 2013.

[152] Yunbo Wang, Mehmet C. Vuran, and S. Goddard. Cross-layer analysis of the end-to-end delay distribution in wireless sensor networks. *IEEE/ACM Transactions on Networking*, 20(1):305–318, Feb 2012.

[153] IEEE Standard for Local and Metropolitan Area Networks−Part 21: Media Independent Handover. *IEEE Std. 802.21-2008*, 2009.

[154] A. Awad, A. Mohamed, and C. F. Chiasserini. Dynamic network selection in heterogeneous wireless networks: A user-centric scheme for improved delivery. *IEEE Consumer Electronics Magazine*, 6(1):53–60, Jan 2017.

[155] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Young-Doo Kim, Eungsun Kim, and Yoon-Chae Cheong. An overview of radio resource management in relay-enhanced OFDMA-based networks. *IEEE Communications Surveys Tutorials*, 12(3):422–438, Third 2010.

[156] Alaa Awad, Amr Mohamed, and Carla Fabiana Chiasserini. User-centric network selection in Multi-RAT systems. *IEEE WCNC workshop on Mobile Edge Computing and IoT*, April 2016.

[157] X. Du, M. Guizani, Y. Xiao, and H. H. Chen. A routing-driven elliptic curve cryptography based key management scheme for heterogeneous sensor networks. *IEEE Transactions on Wireless Communications*, 8(3):1223–1229, March 2009.

[158] X. Du, Y. Xiao, M. Guizani, and H. H. Chen. An effective key management scheme for heterogeneous sensor networks. *Ad Hoc Networks, Elsevier*, 5(1):24–34, Jan 2007.

[159] A. Awad, A. Mohamed, C. Chiasserini, and T. Elfouly. Network association with dynamic pricing over D2D-Enabled heterogeneous networks. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, March 2017.

[160] A. Awad Abdellatif, A. Mohamed, and C. Chiasserini. Concurrent association in heterogeneous networks with underlay D2D communication. In *13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 56–61, June 2017.

[161] K. T. Sweeney, T. E. Ward, and S. F. McLoone. Artifact removal in physiological signals: Practices and possibilities. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):488–500, May 2012.

[162] A. B. said, M. F. Al-Sa'D, M. Tlili, A. A. Abdellatif, A. Mohamed, T. Elfouly, K. Harras, and M. D. O'Connor. A deep learning approach for vital signs compression and energy efficient delivery in mhealth systems. *IEEE Access*, 6:33727–33739, 2018.

Appendix A

List of Publications

A.1 Patent filling

[1] Alaa Awad Abdellatif, A. Mohamed, and C. F. Chiasserini, “Method and System for Data-specific Transceiver Design for efficient IoT Devices," IP disclosure under filling by IPTT.

A.2 Journal and Magazine papers

[1] Alaa Awad Abdellatif, A. Mohamed, C. F. Chiasserini, Mounira Tlili, and Aiman Erbad, “Edge Computing For Smart Health: Context-aware Approaches, Opportunities, and Challenges," Under minor revision in IEEE Network.
[2] Alaa Awad Abdellatif, Ahmed Emam, C. Chiasserini, A. Mohamed, Ali Jaoua, Rabab Ward, “Edge-based Compression and Classification for Smart Healthcare Systems: Concept, Implementation and Evaluation," in Expert Systems with Applications, vol. 117, pp. 1-14, March 2019.
[3] Alaa Awad Abdellatif, A. Mohamed, and C. F. Chiasserini, “User-Centric Networks Selection With Adaptive Data Compression for Smart Health," in IEEE Systems Journal, vol. 12, no. 4, pp. 3618-3628, Dec. 2018.
[4] Alaa Awad Abdellatif, M. G. Khafagy, A. Mohamed, and C. Chiasserini, “EEG-based Transceiver Design with Data Decomposition for Healthcare IoT Applications," in IEEE Internet of Things Journal, vol. 5, no. 5, pp. 3569-3579, Oct. 2018.

[5] Alaa Awad, A. Mohamed, C. F. Chiasserini, and Tarek Elfouly, “Distributed in-network processing and resource optimization over mobile-health systems," Journal of Network and Computer Applications, vol. 82, pp. 65-76, March 2017.
[6] Alaa Awad, A. Mohamed, and C. F. Chiasserini, “Dynamic Network Selection in Heterogeneous Wireless Networks: A user-centric scheme for improved delivery," in IEEE Consumer Electronics magazine, vol. 6, no. 1, pp. 53-60, Jan. 2017.

A.3 Conference papers

[1] Alaa Awad Abdellatif, A. Mohamed, and C. F. Chiasserini, “Automated Class-based Compression for Real-Time Epileptic Seizure Detection," Wireless Telecommunications Symposium (WTS), Arizona, April 2018.
[2] Alaa Awad Abdellatif, A. Mohamed, and C. F. Chiasserini, “Concurrent association in heterogeneous networks with underlay D2D communication," International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, pp. 56-61, June 2017.
[3] Alaa Awad, A. Mohamed, C. F. Chiasserini, and T. Elfouly, “Network Association with Dynamic Pricing over D2D-Enabled Heterogeneous Networks," IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, pp. 1-6, March 2017.
[4] Alaa Awad, A. Saad, A. Jaoua, A. Mohamed and C. F. Chiasserini, “In-Network Data Reduction Approach Based on Smart Sensing," IEEE Global Communications Conference (GLOBECOM), Washington, DC, pp. 1-7, December 2016.
[5] Alaa Awad, A. Mohamed and C. F. Chiasserini, “User-centric network selection in multi-RAT systems," IEEE Wireless Communications and Networking Conference, Doha, pp. 1-6, April 2016.
[6] Alaa Awad, Medhat H. M. Elsayed, A. Mohamed, “Encoding Distortion Modeling For DWT-Based Wireless EEG Monitoring System,” IEEE Annual Consumer Communications & Networking Conference (CCNC), January 2016.
[7] Alaa Awad, A. Mohamed and T. Elfouly, "Energy-cost-distortion optimization for delay-sensitive M-health applications," Wireless Telecommunications Symposium (WTS), New York, NY, pp. 1-5, April 2015.