

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

4/2020

Energy Efficiency in 5G Communications – Conventional to Machine Learning Approaches

M. K. Shahid et al.

Paper

1

Time-shifted Pilot-based Scheduling with Adaptive Optimization for Pilot Contamination Reduction in Massive MIMO

A. P. Kumar and T. Srinivasulu

Paper

10

MIMO Antenna Design and Optimization with Enhanced Bandwidth for Wireless Applications

M. H. Reddy and D. Sheela

Paper

22

A Comprehensive Survey on Resource Management in Internet of Things

L. B. Bhajantri and Gangadharaiah S.

Paper

27

Designing a Compact Microstrip Antenna Using the Machine Learning Approach

K. Sharma and G. P. Pandey

Paper

44

Transmission of Disaster Warnings via Control Channels in Cellular Networks

M. I. Aal-Nouman et al.

Paper

53

Detection of DDoS Attacks in OpenStack-based Private Cloud Using Apache Spark

S. Gumaste, Narayan D. G., S. Shinde, and Amit K

Paper

62

A New Parameterized Model for Determining Quality of Online Video Service Using Modern H.265/HEVC and VP9 Codecs

T. Uhl and C. Hoppe

Paper

72

(Contents Continued on Back Cover)

Editor-in-Chief

Adrian Kliks, *Poznan University of Technology, Poland*

Steering Editor

Jordi Mongay Batalla, *National Institute of Telecommunications, Poland*

Editorial Advisory Board

Hovik Baghdasaryan, *National Polytechnic University of Armenia, Armenia*

Naveen Chilamkurti, *LaTrobe University, Australia*

Luis M. Correia, *Instituto Superior Técnico, Universidade de Lisboa, Portugal*

Luca De Nardis, *DIET Department, University of Rome La Sapienza, Italy*

Nikolaos Dimitriou, *NCSR "Demokritos", Greece*

Ciprian Dobre, *Politechnic University of Bucharest, Romania*

Filip Idzikowski, *Poznan University of Technology, Poland*

Andrzej Jajszczyk, *AGH University of Science and Technology, Poland*

Albert Levi, *Sabancı University, Turkey*

Marian Marciniak, *National Institute of Telecommunications, Poland*

George Mastorakis, *Technological Educational Institute of Crete, Greece*

Constandinos Mavromoustakis, *University of Nicosia, Cyprus*

Klaus Mößner, *Technische Universität Chemnitz, Germany*

Imran Muhammad, *King Saud University, Saudi Arabia*

Mjumo Mzyece, *University of the Witwatersrand, South Africa*

Daniel Negru, *University of Bordeaux, France*

Ewa Orłowska, *National Institute of Telecommunications, Poland*

Jordi Perez-Romero, *UPC, Spain*

Michał Pióro, *Warsaw University of Technology, Poland*

Konstantinos Psannis, *University of Macedonia, Greece*

Salvatore Signorello, *University of Lisboa, Portugal*

Adam Wolisz, *Technische Universität Berlin, Germany*

Tadeusz A. Wysocki, *University of Nebraska, USA*

Publications Staff

Content Editor: **Robert Magdziak**

Managing Editor: **Ewa Kapuściarek**

Technical Editor: **Grażyna Woźnica**

ISSN 1509-4553 on-line: ISSN 1899-8852

© Copyright by National Institute of Telecommunications, Warsaw 2020

Circulation: 300 copies Sowa – Druk na życzenie, www.sowadruk.pl, tel. 22 431-81-40



**Ministry of Science
and Higher Education**
Republic of Poland

Improvement of the linguistic quality of the journal; Assignment of DOI numbers; Access to the antiplagiarism system; Editing and printing of the journal – tasks financed under 706/P-DUN/2019 agreement from budget of the Ministry of Science and Higher Education under the science dissemination fund.

Energy Efficiency in 5G Communications – Conventional to Machine Learning Approaches

Muhammad Khalil Shahid, Filmon Debretsion, Aman Eyob, Irfan Ahmed, and Tarig Faisal

Higher Colleges of Technology, Abu Dhabi, United Arab Emirates

<https://doi.org/10.26636/jtit.2020.146820>

Abstract—Demand for wireless and mobile data is increasing along with development of virtual reality (VR), augmented reality (AR), mixed reality (MR), and extended reality (ER) applications. In order to handle ultra-high data exchange rates while offering low latency levels, fifth generation (5G) networks have been proposed. Energy efficiency is one of the key objectives of 5G networks. The notion is defined as the ratio of throughput and total power consumption, and is measured using the number of transmission bits per Joule. In this paper, we review state-of-the-art techniques ensuring good energy efficiency in 5G wireless networks. We cover the base-station on/off technique, simultaneous wireless information and power transfer, small cells, coexistence of long term evolution (LTE) and 5G, signal processing algorithms, and the latest machine learning techniques. Finally, a comparison of a few recent research papers focusing on energy-efficient hybrid beamforming designs in massive multiple-input multiple-output (MIMO) systems is presented. Results show that machine learning-based designs may replace best performing conventional techniques thanks to a reduced complexity machine learning encoder.

Keywords—5G, energy efficiency, wireless networks.

1. Introduction

Conventional fuels used for power generation, heating and transport have contributed to an 80% increase in greenhouse gas emissions compared to 1970. According to projections pertaining to 2040, global energy demand is expected to increase even further, by 30%, with the pace of growth being even faster in developing countries [1]. 5G networks would inevitably be responsible for an increase in the amount of energy used by consumers, therefore contributing to climate change. As the amount of space within the wave spectrum in which consumer devices may operate is increased by the use of millimeter waves, energy usage grows as well, leading to faster global warming. 3GPP standards, including those related to 5G networks, aim to increase capacity and coverage of the system, with energy efficiency gains considered at architectural and functional level [2]. Ensuring that hardware is capable of working within extended operating condition ranges (temperature and humidity levels prevailing in rooms in which equip-

ment is located) may lead to a decrease in the amounts of power consumed by air conditioning systems. Small cells used to provide 5G connectivity are claimed to be energy efficient and powered in a sustainable way. However, maintenance- and production-related issues may cause considerable cost implications [1]. Deployment of 5G systems is also expected to improve energy efficiency (EE) of the entire industry as a whole, as the cost of energy per bit of data transferred is, in 5G, equal to one tenth of the level experienced in 4G [2]. However, base stations still remain energy-hungry locations of the network, due to the foreseeable increase in traffic that is expected to grow by several thousand percent.

A few papers exist that focus on analyzing EE of 5G networks. Report [3] surveys various optimization techniques, the game theory and machine learning approaches that have been proposed for enhancing power allocation to downlink and uplink channels. Other energy-saving approaches are described therein as well. In paper [1], some of the significant examples discussed include deployment of newer radio resource control (RRC) for context signaling and for reducing the number of redundant state changes. Utilization of advanced clustering and caching techniques on the radio access network (RAN) side has been highly valued for improving latency requested by a group of users and for eliminating the factor of clogging the network by a huge number of requests for the same data. Commercial resource sharing between different operators offers encouraging results in terms of reduced deployment costs and good data rates, while ensuring minimum interference. In a paper [4] a detailed discussion of the various advantages and disadvantages of green and energy efficiency techniques is presented, contributing to understanding the ways in which green radio architecture may be used in 5G and future mobile networks, and presenting the challenges that will be encountered in the process. In this paper, three most promising green solutions are analyzed. Extreme mobile broadband (xMBB) is a service characterized by high data rates, low latency communication (LLC), and extreme coverage. Its spectrum resources include lower bands, and new higher bands with large contiguous bandwidth, (license + LSA + LAA). Its target values are defined by the peak data

rate of up to 20 Gbps for downlink and 10 Gbps for uplink. Massive machine-type communications (mMTC) is another type of service that offers the following features: dense mobile networks, wide-area coverage and deep penetration. It relies on lower band frequencies and its spectrum resources include licensed shared access (LSA) and licensed-assisted access (LAA). The target connectivity density value equals 1 million devices per square kilometer. The third service offers ultra-reliable and low-latency communication links, such as V2X. Its spectrum resources are based on lower bands, exclusive licenses, and its target user latency equals 0.5 ms. A typical 5G network is shown in Fig. 1. Mathematically, energy efficiency of the base station side is defined as:

$$EE = \frac{R}{\eta P_t + P_c} \quad [\text{bits/Joule}], \quad (1)$$

where, R is the average overall data rate in bits per second, η is the reciprocal of the transmit power (amplifier efficiency), P_t is the transmission power, and P_c is total power dissipated in the transmitter circuit.

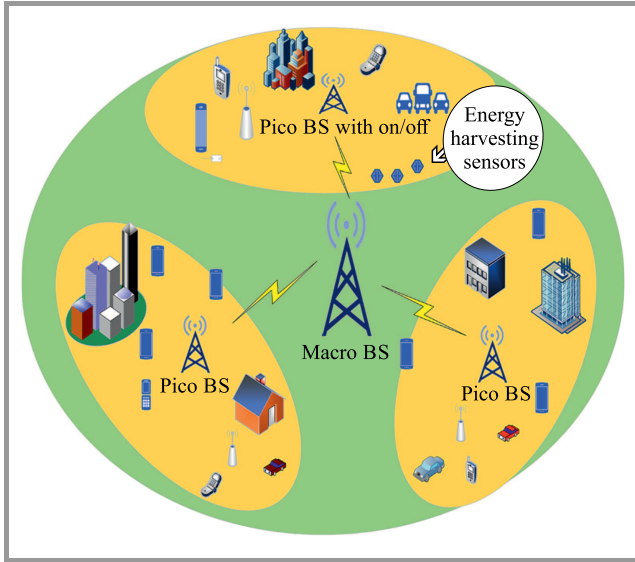


Fig. 1. Model of a multiuser massive MIMO downlink system.

The rest of the paper is organized as follows: in Section 2, on/off techniques are presented and energy harvesting is described. Section 3 presents heterogeneous networks and energy efficiency-related considerations. Energy-efficient physical layer hardware designs are reviewed in Section 4, and machine learning techniques ensuring EE are presented in Section 5.

2. Energy Efficiency Using On/Off Techniques and Energy Harvesting

Authors in [5] propose three independent energy efficiency optimization solutions to minimize energy consumption by either forcing idle base stations to go to sleep, or by dynamically adjusting the signal range of base stations through

software-defined networking. This means that the stream table of the base stations is reconfigured to modify the connections between users and base stations in order to free as high a number of base stations that is feasible under specific circumstances. Secondly, the maximum transmission time required for given content to be downloaded from the server is minimized in order to let the base station to go to rest. Finally, at times a stronger signal may be used by base stations in situations in which weaker signal might be sufficient to cover the needs of all customers. Therefore, the power level of such a station may be adjusted to save energy.

In [6], the authors proposed a strategy for dense femtocell deployment based on sleep mode and hybrid access policies. In most studies, a femtocell base station awakens from rest when its clients are active. However, in this work, the femtocells with dynamic clients may continue to sleep if the reallocation of their clients is feasible based on hybrid access to their neighbors, with throughput enhancements ensured simultaneously. Simulations indicate that an increase in femtocell density boosts the number of clusters formed and femtocells remaining in the sleep mode, hence improving energy efficiency.

Paper [7] describes the efficiency of harvesting energy through radio frequency (RF). Two mathematical simulations are conducted as part of the study, verifying the energy efficiency to training interval ratio and checking how EE is affected by the change in block size and in the number of users. Simulations show that EE is low for large T_t values due to the amount of energy dissipated, which means that the factor that is of most significance for EE deteriorates. Larger amounts of energy may be harvested with more receivers (users). This is because one of the users is decoding information while others harvest energy. In paper [8], the authors propose an iterative hybrid analog-digital beamforming scheme for simultaneous wireless information and power transfer (SWIPT) for MIMO systems with limited RF chains at the base station. Compared to fully-digital SWIPT, the proposed scheme is reported to be a better solution for energy harvesting with significant gains in total power consumption.

3. Energy Efficient Heterogeneous Networks

A potential solution to improve energy efficiency of 5G heterogeneous networks consists in offloading traffic from macro cells to small cells. Paper [9] presents a scheme conditionally offloading traffic from macro cells to picocells and femtocells under the condition that the cell system load of all cells is maintained below a certain threshold. The proposed scheme is developed using an online reinforcement learning methodology capable of surmising that other macro cells are pursuing offloading strategies, which reduces the base station information exchange overhead. The study found that cell load is a key factor that impacts inter-

ference, small cell traffic congestion, and energy efficiency. As network load or cell load increase, energy efficiency, interference, and quality of service (QoS) decrease.

In [10], an energy efficiency comparison has been performed between single-tier, two-tier, and three-tier networks by using stochastic geometry tools. Base stations of all tiers are positioned in accordance with the Poisson point process, and they are not quite the same as one another in terms of their size, density, and transmitting power. Simulation results show that energy efficiency of a heterogeneous network is enhanced by increasing the number of pico and femtocells. However, the growing number of pico and femtocell base stations increases co-channel interference, since pico and femtocells are located inside macro cells.

Multi-connectivity is a feature of 5G technology that allows users to connect and consume resources from multiple base stations concurrently, potentially using different radio interfaces, i.e. Evolved-LTE and new radio. Authors in [11] propose new algorithms for multi-connectivity and compare them with single connectivity scenarios to examine how multi-connectivity is capable of improving reliability and the system's overall energy efficiency. The paper shows that at low speeds, multi-connectivity offers little improvement in terms of radio link failures. However, at high speeds, a reduction of up to 50% is observed. The newly proposed secondary cell selection algorithms presented in [11] improved energy efficiency of the network by up to 20% at low speeds.

The proposal [12] is to optimize power using spectrum sharing in the next generation networks (NGNs) in order to achieve high spectrum and energy efficiency for both the primary and the secondary system, without the involvement of a secondary transmitter. When comparing the performance of the proposed model with the opportunistic spectrum sharing model and other popular resource allocation algorithms, the efficiency of the proposed scheme turns out to be superior. Spectrum sharing is gaining popularity as it allows to expand the spectrum and boost power efficiency of the network. Spectrum sharing allows for cooperative or simultaneous use of limited radio frequency resources by a number of independent users within a particular geographical area. Spectrum sharing may help effectively use white spaces or underutilized portions of the spectrum. Also, there are various power allocation strategies for optimum resource block allocation in the spectrum sharing process. Optimum resource block allocation strategies aim to optimize the spectrum and energy efficiency of the system, simultaneously increasing the quality of service for both primary and secondary receivers.

4. Energy Efficient Physical Layer Hardware Design

Paper [13] aims to reduce both hardware complexity and power consumption in hybrid analog-digital beamforming systems. Reduced-dimension training sequence designs

and transmit precoder designs are considered jointly. Paper [19] discusses the minimum number of RF chains required and trade-offs with the bandwidth of the transmitted signal needed in order to realize any given fully digital precoder relying on hybrid analog/digital precoding (HADP) in wide-band mm-Wave systems. The authors of [10] compare different hybrid beamforming architectures and optimize the number of antennas in order to maximize EE. The combination of phase shifters and switches has been shown to be superior to the conventional phase shifter-only architectures in terms of spectral and energy efficiencies. In the paper [14], the authors propose wireless power transfer (WPT) based algorithms for the hybrid beamforming design to achieve optimized performance of the fully digital beamforming architecture. In HB, the number of RF chains is in general significantly lower than that of transmit antennas. It is shown that for a general point-to-point MIMO WPT system, optimized performance may be achieved as long as the number of RF chains at the energy transmitter is not lower than twice the number of sub-bands used or twice the number of channel paths. For MISO WPT, the required number of RF chains may be reduced even further to equal the number of channel paths only.

Paper [15] compares spectral and energy efficiency of hybrid beamforming and digital beamforming systems, taking into account the effects of channel estimation, transmitter impairments, and multiple simultaneous users for a wideband multipath model. Considering the model of the quantization error at each antenna, better EE and achievable rates are obtained with 5-bit ADC resolution at high SNR.

The digital beamforming model with low resolution ADCs outperforms hybrid beamforming in terms of spectral efficiency and energy efficiency for single-user and multi-user scenarios with multipath propagation [15], [16]. Low resolution ADC DBF is resistant to small automatic gain control (AGC) imperfections. Energy efficiency and spectrum efficiency (SE) of wireless communication systems may be significantly enhanced by large-scale antenna systems. These systems are deployed in hybrid digital and analog BF structures. The analysis of this kind of structures presented in [17] provides an insight into optimal analog and digital BF designs, EE-SE relationship at the green point, and the impact of the number of transceivers on the green EE point, i.e. the point with the highest EE along the EE-SE curve. In paper [18], the authors compare the energy efficiency maximization problem for three different types of precoding scenarios, namely zero-forcing (ZF), general beamforming and conjugate-beamforming, in a multiuser downlink distributed antenna system with a hybrid energy supply. The proposed algorithm, based on joint power allocation and energy cooperation, allows to maximize EE and shows that ZF achieves better EE with low noise variance, while conjugate beamforming performs better in high-noise scenarios. Paper [8] proposes different hybrid beamforming schemes in which phase-only weights are applied for the analog precoder and combiner in a mm-Wave MU-MIMO transmis-

sion. The proposed design is applied in fully-connected structures and is then extended to their sub-connected counterparts. In [19], the authors propose an energy efficient optimization model for multi-user massive MIMO, based on the joint optimization problem involving computation and communication power. An upper bound of energy efficiency is provided for multi-user massive MIMO systems with partially-connected structures of RF transmission systems. The proposed design aims to find an EE suboptimal solution capable of jointly improving computation and communication power in massive MIMO systems. Paper [20] addresses the implications of highly directional communication for the design of an efficient medium access control (MAC) layer. It discusses the key MAC layers options available and design physical control channel aspects that may improve performance, spectral efficiency and energy efficiency. An energy-efficient hybrid beamforming design for a partially-connected structure is investigated in [21]. The authors design a two-layer optimization method to solve the non-convex problem by exploiting interference alignment and fractional programming. First, an analog beamformer is obtained using the alternating-direction optimization method, and then the minimum mean square error (MMSE)-based digital precoder and combiner are formed.

5. Energy Efficiency Using Machine Learning Techniques

In recent research, machine learning techniques are widely explored in various sections of the 5G networks to enhance their performance in terms of spectral and energy efficiency. In [22], the authors propose a centralized resource allocation scheme that exploits online learning techniques, guaranteeing mitigation of interference and maximization of energy efficiency, while avoiding dropping QoS requirements for all users. The priority of users in resource block (RB) allocation and compact state representation-based learning methodology was considered to enhance the learning process in the design with a model-free reinforcement learning. The outcome of the simulation clearly shows that the proposed resource allocation scheme is capable of mitigating interference, considerably increasing energy and spectral efficiency, as well as sustaining the users' QoS requirements. Centralized joint RBs and the power allocation scheme rely on a single controller integrated into the centralized baseband unit. This controller gathers the network state information through its interface with the macro BSs and uses this knowledge to select the most appropriate actions that ensure energy efficiency enhancements, while simultaneously maintaining QoS requirements from different tier users. The proposed online learning model ensures compact state representation in order to reduce the size of the state space, augments the algorithm's convergence and manages the curse of dimensionality. The conventional base station is separated into two parts within a heterogeneous cloud radio access network structure. This allows to

centralize the signal processing part as the baseband unit, and the signal transceiver part as the radio remote head: the centralized baseband unit and the radio remote head unit are associated through fronthaul links. Paper [23] proposes a joint transmission mode selection and power control algorithm with reinforcement learning to ensure energy optimization via fifth-generation communication for the heterogeneous cloud radio access network architecture in a vehicular social network. It considers cellular and D2D communication coexisting in a single cell environment. Two Q-learning algorithms have been designed that enable to take optimal communication mode selection and transmission power control decisions by altering the target SINR. The two Q-learning models are centralized and distributed Q-learning. The agent in the baseband unit uses centralized Q-learning to maximize the system's energy efficiency while guaranteeing compliance with QoS constraints. The vehicles use distributed Q-learning to maximize their achievable data rate.

In [24], a sub-connected switch-based hybrid precoding architecture with lens array for beamspace MIMO systems is proposed to overcome the issue of high energy consumption. The hybrid precoding scheme aiming to maximize the achievable sum rate is based on the cross-entropy (CE) optimization approach developed in machine learning [24]. But this paper does not focus on inter-user interference in the beam domain.

In [25], a machine learning approach was applied to a joint hybrid beamforming and radio resource management for a rank constrained mm-Wave MU-MIMO downlink. Formulation of the optimization problem aimed to maximize the total throughput constraint while taking into account the total transmit power and the number of RF chains. The energy efficiency aspect has not been investigated in this paper. Three different machine learning methods, namely random forest, deep learning, and ridge regression, were applied in [26] to forecast the resulting uplink transmission power based on the passive network quality indicators and application-level information available. The random forest model was reported as offering the best performance. However, this framework cannot be utilized to optimize energy efficiency without the channel feedback overhead. Machine type communications also need a low latency 5G technology. Fast establishment of uplink transmission for machine type communications (MTCs) is one of the main challenges affecting future wireless systems. The optimal selection of machine type devices (MTDs) to be used in the establishment of a fast uplink can be achieved by applying machine learning-based techniques, namely multi-armed bandit theory and deep reinforcement learning [27]. The deep reinforcement learning and multi-armed bandit theory-based uplink resource allocation methodology presented only works for dominant uplink traffic networks, such as sensor networks or MTDs. It cannot be applied to mobile users who require mostly downlink data transmissions. The authors of [28] proposed a genetic algorithm (GA) approach to resource block allocation in a multi-cell network. This

approach allows to minimize inter-cell interference by maximizing channel capacity in the resource allocation process. The proposed genetic algorithm-based resource allocation methods for cell-edge users uses the multipath Rayleigh channel. However, 5G communication is supposed to rely on the mm-Wave channel, with dominant line-of-sight (LOS) transmission paths. The resource allocation problem of small LTE-U base stations (SBSs) was modeled based on the deep learning approach in [29]. Dynamic channel selection, carrier aggregation, and fractional spectrum access operations are performed proactively, while guaranteeing fairness with existing Wi-Fi networks and other LTE-U operators. Again, this work does not consider mm-Wave channels and the proposed solution is intended for the multipath channel model.

The author of [30] proposed an actor-critic reinforcement learning (RL) approach based on the stochastic policy gradient to improve the system's throughput and to boost D2D throughput. Compared to the value-based scheme, such as Q-learning, the policy-based method was reported to be better, since it maximized the expected throughput by searching in the policy space. Due to the near-optimal and direction beam transmission in the mm-Wave channel, D2D communication will exert lower impact on the existing cellular network. Reinforcement learning has also been proposed for the resource allocation stage in mobile edge computing systems [31]. It offers a near-optimal solution for the joint task of offloading and resource allocation, simultaneously minimizing energy consumption and delays. This work does not consider transmission power consumption. In [32], the resource allocation problem in virtual reality wireless networks is considered for both uplink and downlink scenarios. Formulation of the has led to the establishment of a non-cooperative game and a distributed algorithm based on the machine learning framework of echo state networks (ESNs) being proposed to solve this game. Such a model is used for determining VR metrics, such as tracking accuracy, processing delay, and transmission delay. Deep reinforcement learning was applied to develop a decentralized resource allocation mechanism with minimum transmission overhead for vehicle-to-vehicle (V2V) communications [33]. The decisions concerned with finding the optimal sub-band and power level for the transmission are autonomously taken to support each V2V link. This approach does not use massive MIMO or hybrid beamforming. In [34], the authors proposed a simple reinforcement learning algorithm based on an epsilon-greedy multi-hop to achieve significant energy savings in low-power wide area networks relying on multi-hop topologies. In single-hop topologies, uplink (UL) communications from distant nodes are established with high power levels, whereas transmissions to closer hops are enabled by multi-hop routing in UL, in order to reduce energy consumption. This LoWAN network uses remote with IEEE 802.15.4 physical and MAC layers. The proposed multi-hop dynamic routing approach is not always energy efficient. There is a tradeoff between circuit power consumption and transmission power. There

is a need for cross-layer machine learning techniques covering PHY, MAC, and network layers to ensure optimal EE and SE levels.

6. Performance Study

In this section, we summarize our preliminary studies concerning machine learning techniques deployed for user-beam pair selection in the hybrid beamforming design used in 5G communications. Hybrid beamforming is an imperative part of 5G massive MIMO millimeter wave communications [35]. Hybrid beamforming divides the processing of signal for beamforming into two stages: analog beamforming and digital beamforming. Analog beamforming is realized by analog phase shifters, while digital beamforming usually relies on zero-forcing precoders/combiners. This arrangement reduces the number of radio frequency chains required in the digital beamforming and reduces energy consumption. Energy consumption may be further reduced by user-beam allocation in the analog beam domain.

6.1. Hybrid Beamforming and User-Beam Selection Architecture

In hybrid beamforming, beamforming is performed in two stages: an analog beamforming $\mathbf{F}_{AB} \in \mathbb{C}^{N \times N_{RF}}$, and digital beamforming $\mathbf{F}_{DB} \in \mathbb{C}^{N_{RF} \times N_s}$ as shown in Fig. 2.

We consider a single base station scenario. The base station is equipped with $N = 128$ transmit antennas and is serving $K = 8$ users. There are $N_{RF} = K$ RF chains available within the base station. We use an extended Saleh-Valenzuela model which accurately captures the mathematical structure of in mm-Wave channels [36]. For simplicity, we assume that each scattering cluster around the transmitter and the receiver contributes a single propagation path [37]. The near optical line-of-sight (LOS) wave propagation at mm-Wave frequencies results in a limited number of scattering paths, say L . The $N \times K$ MIMO channel matrix at the BS can be written as:

$$\mathbf{H} = \sqrt{\frac{KN}{L}} \sum_{l=1}^L \alpha_l \mathbf{a}(\phi_l), \quad (2)$$

where α_l represents the complex gain of the l -th path with i.i.d. $\mathcal{CN}(0, 1)$. Moreover, $\mathbf{a}$ is the transmit steering vector. The variable $\phi_l \in [0, 2\pi)$ is the l -th path's azimuth angle.

The analog beamforming matrix, e.g. discrete Fourier transform (DFT) $N \times N$ matrix $\mathbf{U}$ may be used to convert spatial channel $\mathbf{H}$ to a beam domain channel as:

$$\mathbf{H}_b = \mathbf{UH}. \quad (3)$$

Now, the row of beam domain channel $\mathbf{H}_b$ corresponds to N beams. The selection of K beams out of N to maximize the energy efficiency poses a complex user-beam matching

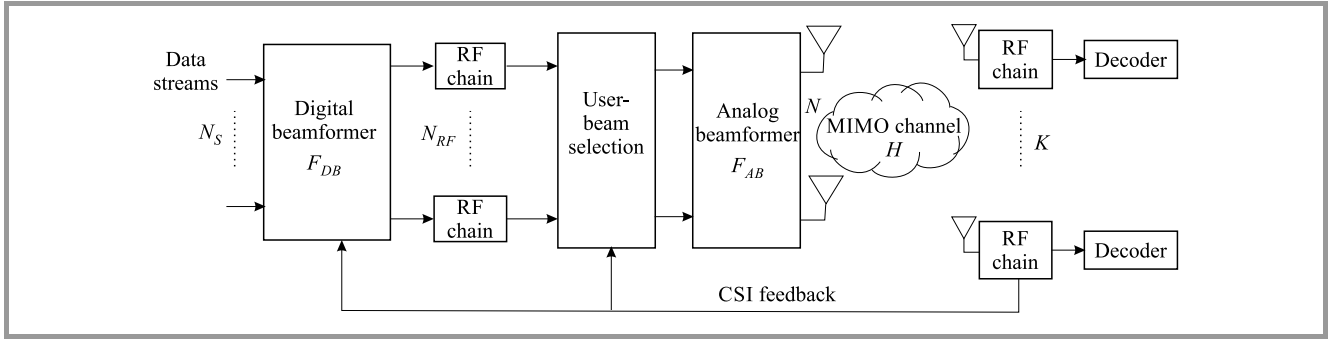


Fig. 2. Hybrid beamforming and user-beam selection architecture.

problem. User-beam allocation is an NP-hard mixed integer programming optimization problem. A global optimal solution to this problem with $K \leq N_{RF}$ (K is the number of users and N_{RF} is the number of RF chains) may only be obtained by using an exhaustive search method. This solution has very high computational complexity of:

$$\sum_{i=1}^{N_{RF}} K C_i \cdot {}^N C_i \cdot P_i,$$

where ${}^a C_b = \frac{a!}{b!(a-b)!}$ and ${}^a P_b = \frac{a!}{(a-b)!}$, $a \geq b$.

For the data symbol vector $\mathbf{s} \in \mathbb{C}^{K \times 1}$, the beam domain representation of the system is given by:

$$\mathbf{y} = \mathbf{H}_b^H \mathbf{F}_{DB,b} \mathbf{s} + \mathbf{z}, \quad (4)$$

where $\mathbf{F}_{DB,b} = \mathbf{U}^H \mathbf{F}_{DB} \in \mathbb{C}^{M \times K}$ is the beam domain digital precoder and $\mathbf{z} \in \mathbb{C}^{N \times 1}$ is the independent and identically distributed (i.i.d.) complex Gaussian noise with zero-mean and variance of one, i.e. $\mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$. The signal received at user k is given as:

$$y_k = \mathbf{h}_{k,b}^H \mathbf{f}_{DB,k,b} \mathbf{s}[n] + z_k. \quad (5)$$

The received SNR γ_k of the user k is:

$$\gamma_k = \frac{|\tilde{\mathbf{h}}_{k,b}^H \mathbf{f}_{DB,k,b}|^2}{\sum_{j \neq k} |\tilde{\mathbf{h}}_{j,b}^H \mathbf{f}_{DB,k,b}|^2 + \frac{1}{\rho} \|\mathbf{f}_{DB,k,b}\|_2^2}, \quad (6)$$

where ρ is the average received SNR per antenna per user and $\tilde{\mathbf{h}}_{k,b}$ is the beam domain channel after the user-beam selection. The sum-rate is given as:

$$R = \sum_{k=1}^K \log_2(1 + \gamma_k). \quad (7)$$

Finally, the energy efficiency is calculated by Eq. (1).

In our machine learning model, training data is obtained from the Hungarian method [38] and is then applied to the user-beam matching problem. We use a shallow neural network with the input layer of 128 neurons, the hidden layer of 20 neurons and the output layer of 8 neurons. The

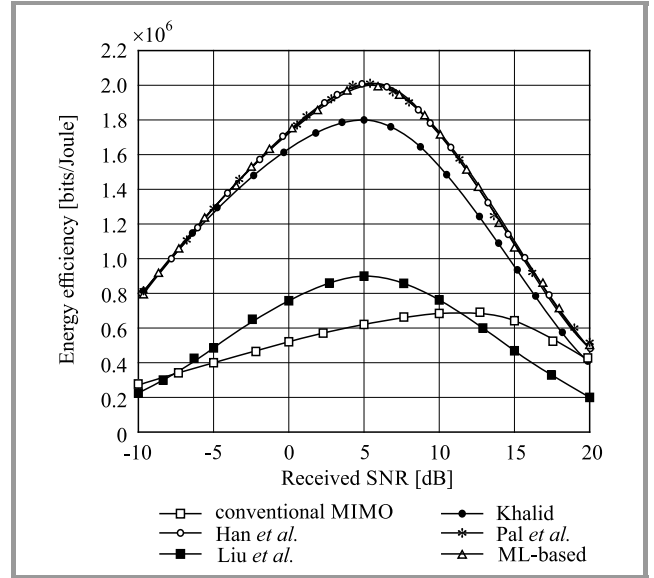


Fig. 3. Energy efficiency versus received SNR with 8 users and 128 transmit antennas.

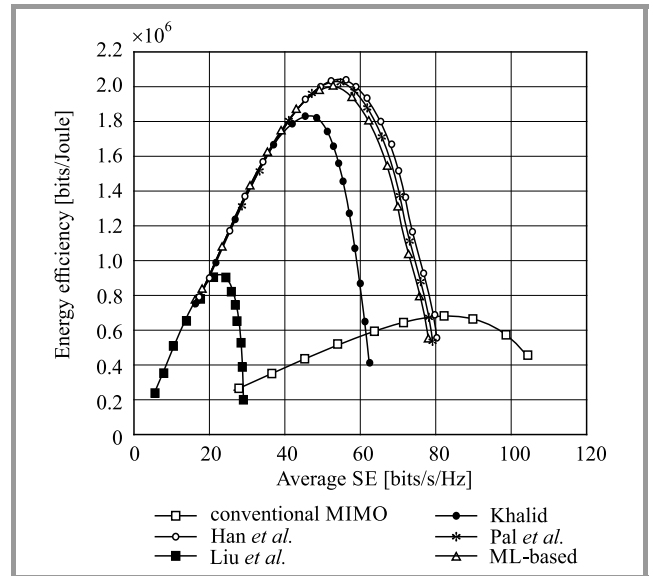


Fig. 4. Energy efficiency and spectral efficiency trade-off with 8 users and 128 transmit antennas.

Table 1
Energy efficiency comparison of some recent research works devoted to signal processing

Authors	Reference paper	Energy efficiency
Han <i>et al.</i>	[39]	High
Liu <i>et al.</i>	[41]	Low
Khalid	[42]	Medium
Pal <i>et al.</i>	[40]	high
ML-based	This paper	High
Conventional MIMO	Benchmark	Low

hidden and output layers use sigmoid activation functions. Figure 3 shows the average system level EE for the received SNR. EE increases with the received SNR, up to the level of 5 dB, and then starts decreasing due to interference and high transmit power. ML-based hybrid beamforming offers performance that is comparable to that of DFT-based [39] and DFT+Greedy-based [40] solutions, but required fewer computational resources (Table 1). It may be observed that conventional MIMO precoding offers the lowest EE performance because of $N_{RF} = N$ RF chains, while in other hybrid beamforming designs $N_{RF} = K$. The trade-off between EE and spectral efficiency (SE) is shown in Fig. 4. EE increases with SE up to a certain level, and then EE starts decreasing. ML-based hybrid beamforming may be used to reduce the computational complexity and achieve comparable, good performance.

7. Conclusions

This article reviews energy efficiency of 5G wireless networks relying on conventional and machine learning techniques. Energy efficiency may be improved by deploying the following techniques: base station on/off, energy harvesting with simultaneous transfer of wireless information and power, small cells, coexistence of LTE and 5G and signal processing algorithms. The latest machine learning techniques may be applied as well. Having analyzed the use of various energy efficiency techniques in 5G networks, we present a comparison of different hybrid beamforming designs in mm-Wave massive MIMO systems. It has been shown that the two-layered shallow feedforward neural network-based ML scheme provides a negligible performance gap. It is expected that machine learning will replace adaptive and computationally intensive portions of future communication systems.

Acknowledgments

This work was supported under HCT Research Grants, grant number 2307 from Higher Colleges of Technology (HCT), Abu Dhabi, United Arab Emirates.

References

- [1] “What Will 5G Mean for the Environment?”, The Henry M. Jackson School of International Studies”, University of Washington [Online]. Available: <https://jsis.washington.edu/news/what-will-5g-mean-for-the-environment/>
- [2] “Power a Greener 5G Era with Huawei 5G Power”, Mobile World Live” [Online]. Available: <https://www.mobileworldlive.com/huawei-updates/power-a-greener-5g-era-with-huawei-5g-power/>
- [3] M. Usama and M. Erol-Kantarci, “A survey on recent trends and open issues in energy efficiency of 5G”, *Sensors* (Basel), vol. 19, no. 14, 2019 (DOI: 10.3390/s19143126).
- [4] A. Bohli and R. Bouallegue, “How to meet increased capacities by future green 5G networks: a survey”, *IEEE Access*, vol. 7, pp. 42220–42237, 2019 (DOI: 10.1109/ACCESS.2019.2907284).
- [5] Q. Zhao and M. Gerla, “Energy efficiency enhancement in 5G mobile wireless networks”, in *Proc. 2019 IEEE 20th IEEE Int. Symp. on a World of Wirel., Mob. and Multim. Netw. WoWMoM 2019*, Washington, DC, USA, 2019 (DOI: 10.1109/WoWMoM.2019.8792998).
- [6] C. Bouras and G. Diles, “Energy efficiency in sleep mode for 5G femtocells”, in *Proc. 2017 Wireless Days WD 2017*, Porto, Portugal, 2017, no. 3, pp. 143–145 (DOI: 10.1109/WD.2017.7918130).
- [7] F. Mukhlif, K. A. B. Nooridin, Y. A. Al-Gumaei, and A. S. Al-Rassas, “Energy harvesting for efficient 5G networks”, in *Proc. 2018 Int. Conf. on Smart Comput. and Electron. Enterp. ICSCEE 2018*, Shah Alam, Malaysia, 2018 (DOI: 10.1109/ICSCEE.2018.8538376).
- [8] A. Li and C. Masouros, “Energy efficient MIMO SWIPT by hybrid analog-digital beamforming”, in *Proc. 2017 IEEE 18th Worksh. on Sig. Process. Adv. in Wirel. Commun. SPAWC 2017*, Sapporo, Japan, 2017 (DOI: 10.1109/SPAWC.2017.8227793).
- [9] I. Alqerm and B. Shihada, “Energy efficient traffic offloading in multi-tier heterogeneous 5G networks using intuitive online reinforcement learning”, *IEEE Trans. on Green Commun. and Network.*, vol. 3, no. 3, pp. 691–702, 2019 (DOI: 10.1109/TGCN.2019.2916900).
- [10] S. Rizvi *et al.*, “An investigation of energy efficiency in 5G wireless networks”, in *Proc. 2017 Int. Conf. on Circ., Syst. and Simul. ICCSS 2017*, London, UK, 2017, pp. 142–145 (DOI: 10.1109/CIRSYSSIM.2017.8023199).
- [11] V. Poirot, M. Ericson, M. Nordberg, and K. Andersson, “Energy efficient multi-connectivity algorithms for ultra-dense 5G networks”, *Wirel. Netw.*, vol. 26, no. 3, pp. 2207–2222, 2020 (DOI: 10.1007/s11276-019-02056-w).
- [12] H. Kour and R. K. Jha, “Power optimization using spectrum sharing for next generation cellular networks”, in *Proc. 2019 11th Int. Conf. on Commun. Syst. and Netw. COMSNETS 2019*, Bengaluru, India, 2019, vol. 2061, pp. 562–564 (DOI: 10.1109/COMSNETS.2019.8711197).
- [13] S. Noh *et al.*, “Training sequence design for feedback assisted hybrid beamforming in massive MIMO systems”, *IEEE Trans. on Commun.*, vol. 64, no. 1, pp. 187–200, 2016 (DOI: 10.1109/TCOMM.2015.2498184).
- [14] L. Yang, Y. Zeng, and R. Zhang, “Wireless power transfer with hybrid beamforming: How many RF chains do we need?”, *IEEE Trans. on Wirel. Commun.*, vol. 17, no. 10, pp. 6972–6984, 2018 (DOI: 10.1109/TWC.2018.2865313).
- [15] K. Roth, H. Pirzadeh, A. Lee Swindlehurst, and J. A. Nossek, “A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect CSI”, *IEEE J. on Selec. Topics in Sig. Process.*, vol. 12, no. 3, pp. 484–498, 2018 (DOI: 10.1109/JSTSP.2018.2813973).
- [16] K. Roth and J. A. Nossek, “Achievable rate and energy efficiency of hybrid and digital beamforming receivers with low resolution ADC”, *IEEE J. on Selec. Areas in Commun.*, vol. 35, no. 9, pp. 2056–2068, 2017 (DOI: 10.1109/JSAC.2017.2720398).
- [17] S. Han, C. L. Io, C. Rowell, Z. Xu, S. Wang, and Z. Pan, “Large scale antenna system with hybrid digital and analog beamforming structure”, in *Proc. IEEE Int. Conf. on Commun. Worksh. ICC 2014*, Sydney, NSW, Australia, 2014, pp. 842–847 (DOI: 10.1109/ICCW.2014.6881305).

- [18] P. Zhu, H. Mao, J. Li, and X. You, "Energy efficient joint energy cooperation and power allocation in multiuser distributed antenna systems with hybrid energy supply", *IET Commun.*, vol. 13, no. 2, pp. 153–161, 2019 (DOI: 10.1049/iet-com.2018.5181).
- [19] X. Ge, Y. Sun, H. Gharavi, and J. Thompson, "Joint optimization of computation and communication power in multi-user massive MIMO systems", *IEEE Trans. on Wirel. Commun.*, vol. 17, no. 6, pp. 4051–4063, 2018 (DOI: 10.1109/TWC.2018.2819653).
- [20] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective", *IEEE Trans. on Commun.*, vol. 63, no. 10, pp. 3437–3458, 2015 (DOI: 10.1109/TCOMM.2015.2456093).
- [21] S. He, C. Qi, Y. Wu, and Y. Huang, "Energy-efficient transceiver design for hybrid sub-array architecture MIMO systems", *IEEE Access*, vol. 4, pp. 9895–9905, 2016 (DOI: 10.1109/ACCESS.2017.2649539).
- [22] I. Al Qerm and B. Shihada, "Enhanced machine learning scheme for energy efficient resource allocation in 5G heterogeneous cloud radio access networks", in *Proc. IEEE Int. Symp. on Pers., Indoor and Mob. Radio Commun. PIMRC 2018*, Montreal, QC, Canada, 2018, vol. 2017-October (DOI: 10.1109/PIMRC.2017.8292227).
- [23] H. Park and Y. Lim, "Reinforcement learning for energy optimization with 5g communications in vehicular social networks", *Sensors (Basel)*, vol. 20, no. 8, 2020 (DOI: 10.3390/s20082361).
- [24] T. Ding, Y. Zhao, L. Li, D. Hu, and L. Zhang, "Hybrid precoding for beamspace MIMO systems with sub-connected switches: a machine learning approach", *IEEE Access*, vol. 7, pp. 143273–143281, 2019 (DOI: 10.1109/ACCESS.2019.2944061).
- [25] I. Ahmed and H. Khammari, "Joint machine learning based resource allocation and hybrid beamforming design for massive MIMO systems", in *Proc. 2018 IEEE Globecom Worksh. GC Wkshps 2018*, Abu Dhabi, United Arab Emirates, 2019 (DOI: 10.1109/GLOCOMW.2018.8644454).
- [26] R. Falkenberg, B. Sliwa, N. Piatkowski, and C. Wietfeld, "Machine learning based uplink transmission power prediction for LTE and upcoming 5G networks using passive downlink indicators", in *Proc. 2018 IEEE 88th Veh. Technol. Conf. VTC-Fall 2019*, Chicago, IL, USA, 2019, vol. 2018-August (DOI: 10.1109/VTCFall.2018.8690629).
- [27] S. Ali, N. Rajatheva, and W. Saad, "Fast uplink grant for machine type communications: challenges and opportunities", *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 97–103, 2019 (DOI: 10.1109/MCOM.2019.1800475).
- [28] S. Essassi, S. Cherif, and M. Siala, "RB allocation based on genetic algorithm and coordination over the $\times 2$ interface in the LTE uplink", in *Proc. 2013 IEEE 24th Int. Symp. on Pers., Indoor and Mob. Radio Commun. PIMRC 2013*, London, UK, 2013, pp. 2424–2428 (DOI: 10.1109/PIMRC.2013.6666552).
- [29] U. Challita, L. Dong, and W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective", *IEEE Trans. on Wirel. Commun.*, vol. 17, no. 7, pp. 4674–4689, 2018 (DOI: 10.1109/TWC.2018.2829773).
- [30] P. Khuntia and R. Hazra, "An actor-critic reinforcement learning for device-to-device communication underlying cellular network", in *Proc. 2018 IEEE Region 10 Ann. Int. Conf. TENCON 2019*, Jeju, South Korea, 2019, vol. 2018-October, pp. 50–55 (DOI: 10.1109/TENCON.2018.8650160).
- [31] L. Huang, X. Feng, C. Zhang, L. Qian, and Y. Wu, "Deep reinforcement learning-based joint task offloading and bandwidth allocation for multi-user mobile edge computing", *Digit. Commun. and Netw.*, vol. 5, no. 1, pp. 10–17, 2019 (DOI: 10.1016/j.dcan.2018.10.003).
- [32] M. Chen, W. Saad, and C. Yin, "Virtual reality over wireless networks: Quality-of-service model and learning-based resource management", *IEEE Trans. on Commun.*, vol. 66, no. 11, pp. 5621–5635, 2018 (DOI: 10.1109/TCO).
- [33] H. Ye and G. Y. Li, "Deep reinforcement learning for resource allocation in V2V communications", in *Proc. IEEE Int. Conf. on Commun. ICC 2018*, Kansas City, MO, USA, 2018, vol. 2018-May (DOI: 10.1109/ICC.2018.8422586).
- [34] S. Barrachina-Munoz, T. Adame, A. Bel, and B. Bellalta, "Towards energy efficient LPWANs through learning-based multi-hop routing", in *Proc. IEEE 5th World Forum on Internet of Things WF-IoT 2019*, Limerick, Ireland, 2019, pp. 644–649 (DOI: 10.1109/WF-IoT.2019.8767193).
- [35] I. Ahmed *et al.*, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives", *IEEE Commun. Surv. and Tutor.*, vol. 20, no. 4, pp. 3060–3097, 2018 (DOI: 10.1109/COMST.2018.2843719).
- [36] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale MIMO systems", in *Proc. IEEE Int. Conf. on Acoust., Speech and Sig. Process. CASSP 2015*, Brisbane, QLD, Australia, 2015, pp. 2929–2933 (DOI: 10.1109/ICASSP.2015.7178507).
- [37] O. E. Ayach, R. W. Heath, S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems", in *Proc. IEEE Int. Conf. on Commun. ICC 2012*, Ottawa, ON, Canada, 2012, pp. 3724–3729 (DOI: 10.1109/ICC.2012.6363634).
- [38] H. W. Kuhn, "The Hungarian method for the assignment problem", *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: <https://onlinelibrary.wiley.com/doi/full/10.1002/nav.3800020109> (DOI: 10.1002/nav.3800020109).
- [39] Y. Han, S. Jin, J. Zhang, J. Zhang, and K. K. Wong, "DFT-based hybrid beamforming multiuser systems: Rate analysis and beam selection", *IEEE J. of Sel. Topics in Sig. Process.*, vol. 12, no. 3, pp. 514–528, 2018 (DOI: 10.1109/JSTSP.2018.2821104).
- [40] R. Pal, A. K. Chaitanya, and K. V. Srinivas, "Low-complexity beam selection algorithms for millimeter wave beamspace MIMO systems", *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 768–771, 2019 (DOI: 10.1109/LCOMM.2019.2902147).
- [41] Y. Liu and J. Wang, "Low-complexity OFDM-based hybrid precoding for multiuser massive MIMO systems", *IEEE Wirel. Commun. Lett.*, vol. 9, no. 8, pp. 263–266, 2020 (DOI: 10.1109/LWC.2019.2929518).
- [42] F. Khalid, "Hybrid beamforming for millimeter wave massive multiuser MIMO systems using regularized channel diagonalization", *IEEE Wirel. Commun. Lett.*, vol. 8, no. 3, pp. 705–708, 2019 (DOI: 10.1109/LWC.2018.2886882).



Muhammad Khalil Shahid received his B.Sc. in Electrical Engineering from UET Lahore, Pakistan in 1998 and M.Sc. in Electrical Engineering in 2005. He got his Ph.D. in Engineering from BUPT, China in 2008. He has approximately 15 years of hands-on experience, working as an engineer and consultant in the telecom industry. He also has more than 10 years of experience as a trainer and faculty member. His areas of research include wireless communication and networks, optical communication and networks, ICT, as well as telecom policies and regulations. Currently, he is working as Program Team Leader/Assistant Professor at Higher Colleges of Technology, Al-Dhafra Region, UAE.

E-mail: kshahid@hct.ac.ae
Faculty of Electrical Engineering
Higher Colleges of Technology
RUWC/MZWC
Abu Dhabi, United Arab Emirates



Filmon Debretsion is currently an undergraduate student at the Electrical Engineering Department of Higher Colleges of Technology, UAE.

Higher Colleges of Technology
Abu Dhabi, United Arab Emirates



Aman Eyob Abraham is currently an undergraduate student at the Electrical Engineering Department of Higher Colleges of Technology, UAE.

Higher Colleges of Technology
Abu Dhabi, United Arab Emirates



Irfan Ahmed received his B.E. in Electrical Engineering degree and the M.Sc. in Computer Engineering degree from the University of Engineering and Technology, Taxila, Pakistan, in 1999 and 2003, respectively, and the Ph.D. degree in Telecommunication Engineering from the Beijing Uni-

versity of Posts and Telecommunications, Beijing, China, in 2008. Currently, he is working as Associate Professor at Higher Colleges of Technology, UAE. Between 2011 and 2017, he worked as Assistant/Associate Professor in Taif University, KSA. He was post-doctoral fellow with Qatar University from 2010 to 2011, where he worked on two research projects: wireless mesh networks in cooperation with Purdue University, USA, and radio resource allocation for LTE, with Qtel. He was also involved in a National ICT Pakistan-funded research project titled “Design and development of MIMO and Cooperative MIMO test-bed” at Iqra University, Islamabad, Pakistan, from 2008 to 2010. His research interests include wireless LAN (WLAN), medium access control (MAC) protocol design and analysis, cooperative communications, MIMO communications, performance analysis of wireless channels, energy constrained wireless networks, cognitive radio networks, and radio resource allocation. He is the author of more than 25 international publications.

E-mail: irfanahmed44@gmail.com
Higher Colleges of Technology
Abu Dhabi, United Arab Emirates



Tarig Faisal received his Ph.D. degree in Signal Processing from University of Malaya, Malaysia, in 2011. His research focuses on signal processing, intelligent systems, robotics and control. He has almost 20 years of academic and industry experience. Currently, he is the Dean of Academic Operations at Higher Colleges of

Technology.
E-mail: tfaisal@hct.ac.ae
Higher Colleges of Technology
Abu Dhabi, United Arab Emirates

Time-shifted Pilot-based Scheduling with Adaptive Optimization for Pilot Contamination Reduction in Massive MIMO

Ambala Pradeep Kumar and Tadisetty Srinivasulu

KU College of Engineering and Technology, Kakatiya University, Warangal, Telangana, India

<https://doi.org/10.26636/jtit.2020.145020>

Abstract—Massive multiple-input multiple-output (MIMO) is considered to be an emerging technique in wireless communication systems, as it offers the ability to boost channel capacity and spectral efficiency. However, a massive MIMO system requires huge base station (BS) antennas to handle users and suffers from inter-cell interference that leads to pilot contamination. To cope with this, time-shifted pilots are devised for avoiding interference between cells, by rearranging the order of transmitting pilots in different cells. In this paper, an adaptive-elephant-based spider monkey optimization (adaptive ESMO) mechanism is employed for time-shifted optimal pilot scheduling in a massive MIMO system. Here, user grouping is performed with the sparse fuzzy c-means (Sparse FCM) algorithm, grouping users based on such parameters as large-scale fading factor, SINR, and user distance. Here, the user grouping approach prevents inappropriate grouping of users, thus enabling effective grouping, even under the worst conditions in which the channel operates. Finally, optimal time-shifted scheduling of the pilot is performed using the proposed adaptive ESMO concept designed by incorporating adaptive tuning parameters. The efficiency of the adaptive ESMO approach is evaluated and reveals superior performance with the highest achievable uplink rate of 43.084 bps/Hz, the highest SINR of 132.9 dB, and maximum throughput of 2.633 Mbps.

Keywords—massive MIMO, pilot contamination, sparse FCM, time-shifted pilot scheduling, user grouping.

1. Introduction

The advancement of big data and the demand for communication networks has increased considerably, meaning that classical wireless networks, such as 4G, are not able to meet the demand of network users [1]. Massive MIMO is a fundamental technique for establishing 5G wireless communications and maximizing spectral efficiency. It offers the ability to deploy different communication models, like cooperative communication, orthogonal frequency division multiplexing (OFDM), and multicarrier commu-

nication. Massive MIMO relies on varying frequencies in the frequency division duplex (FDD) mode. Thus, massive MIMO is based on the time division duplex (TDD) system [2], which utilizes channel reciprocity for obtaining the channel state information (CSI) [3] required. However, as anticipated, the lower coherence interval limits the count of orthogonal pilots assigned to a user. The use of similar pilots by various cell users means that pilot contaminations cannot be distinguished at the base station (BS). Furthermore, massive MIMO has acquired much interest because of gains in energy and spectrum efficiency affecting BS with huge antennas that serve a set of users concurrently [4]. Under multi-cell scenarios, inevitable reuse of pilots in varying cells is experienced and channel estimates generated from certain cells are prejudiced by the reused pilots which are then sent by users from other cells. The phenomenon is termed as pilot contamination [5].

Despite its potential, the massive MIMO model poses various limitations. One of them is the effect of pilot contamination. The problem occurs when users of adjoining cells utilize similar sequences of pilots. The most unpleasant cases are supposed to occur when neighboring cells transmit similar sequences of a pilot at a similar time. For avoiding the issue of synchronized pilots between adjoining cells, a time-shifted pilot method is devised in which some of the cells transmits downlink data, whereas others send pilots. This technique enhances transmission rates of the system. However, the assessment is based on the supposition that the number of antennas at BS is unlimited. Due to the infinite number of antennas, some interference may be considered to be negligible [6]. Pilot contamination unavoidably limits the efficiency of a huge-scale antenna of a massive MIMO model, because of the corrupted estimation of channels. The recycling of frequency amongst adjacent cells led to interference that limited the quality of service offer to cellular users, especially those positioned at cell edges. Service providers look for solutions to restore performance at low SINR cell locations. Numerous methods are devised to mitigate inter-cell interference.

Among these are solutions that rely on multiple antennas, meaning that on the BS side, these arrays become more affordable [7].

In the literature, various methods are devised to mitigate the effects of pilot contamination. The greedy pilot allocation technique allows to alleviate the effect of contaminating pilots by using arithmetical channel covariance information, but still suffers from a number of complications. A blind technique using subspace partition [8] is used for minimizing inter-cell interference (ICI), wherein channel vectors of different users are supposed orthogonal. An adaptive pilot clustering technique [9] is devised based on coalitional game theory. It may alleviate pilot contamination by employing a subset of pilot resources in an individual cell, thus causing spectral efficiency loss. In [10], the pilot contamination precoding (PCP) method is devised, wherein a precoding matrix is using zero-forcing (ZF) for obtaining an infinite signal-to-interference plus noise ratio (SINR), but it affects entire systems. In [11], an optimal technique for determining a simple suboptimal algorithm and an optimum pilot contamination precoding matrix are devised. In paper [12], an intelligent pilot allocation method is devised that could increase uplink SINR of every user in a target cell with large-scale fading. The pilot power control method shown in [13], [14] is devised for classified cells and allows to minimize pilot contamination, improve downlink reachability and foster efficiency of the entire system.

In [15], a ZF time-shifted pilot scheme is utilized for alleviating pilot contamination by using conjugate beamforming. In [16], the interference cancelation (IC) precoding method is devised for mitigating pilot contamination in massive MIMO models. The method improves service quality and allows for effective user scheduling. In [17], a pilot contamination reduction method is devised that is based on complicated exponentials. In this paper, linear time-varying (LTV) channels are evaluated with optimum pilot symbols and they are devised using minimum mean square error (MMSE) criteria. It is proved that the optimum pilot method is capable of setting successive pilot tones and allocating all pilot clusters uniformly within the frequency domain.

The goal is to use time-shifted pilots for mitigating pilot contamination, and thereby assuring effective pilot scheduling in massive MIMO. The motive of the proposed scheduling technique is to utilize time-shifted pilots to initiate pilot scheduling. Initially, user grouping is performed based on an algorithm that gather users as center and edge user groups, depending on various pilot contamination levels. The grouping of users is performed using sparse FCM, based on attributes that involve large-scale fading factors, SINR and user distance in such a way that inappropriate grouping of users is prevented, enabling effective grouping even under the worst channel conditions. Then, pilot scheduling is performed based on the proposed adaptive ESMO algorithm designed by incorporating adaptive tuning parameters in the ESMO algorithm.

The major contribution of the research performed is to devise an approach enabling time-shifted optimal scheduling of pilots using the adaptive ESMO algorithm that is designed by combining adaptive tuning parameters with ESMO.

The remaining sections of the paper are arranged in the following manner: Section 2 elaborates on the description of conventional pilot scheduling strategies and on the challenges faced. The model of a massive MIMO system is illustrated in Section 3. The proposed method for time-shifted pilot scheduling is portrayed in Section 4. Outcomes of the proposed strategy and its comparisons with other methods are depicted in Section 5. Section 6 presents the conclusions.

2. Literature Survey

Akgun *et al.* devised, in [18], an attack model in which the foe infected the uplink pilot transmissions of multiple users, i.e. they conducted a pilot contamination attack. The downlink transmission rates with/without the attack were derived by exploiting the channel hardening effect of massive MIMO. This model did not consider attack models that involved multiple frequencies and time slots. Wu *et al.* in [19] devised secure communication for the time duplex MIMO model, considering eavesdropping. The eavesdropper attacked the transmission of the uplink pilot before eavesdropping on downlink transmissions. Moreover, pilot signals and data signals were utilized for estimating the uplink channel. This method maximized the asymptotic estimation error and suffered from loss of performance. Fan *et al.* in [20] devised a fractional pilot reuse (FPR) mechanism for minimizing the contamination of pilots from the massive MIMO model. In FPR, users are split into the center and edges of the cell. Then, cell users utilize orthogonal pilots for preventing interference caused by adjoining cells, meanwhile the cell center reutilizes similar pilots for enhancing the efficiency of pilots. The throughput of the network was influenced by the count of service users and pilots.

Saraereh *et al.* in [21] presented a method deployed in a massive MIMO model for decontaminating the set of pilot sequences. The first step is based on path loss to perform grouping, and the second step is based on the pseudo-random code. However, the method failed to consider pilot contamination in sectoral cells. Shaalan *et al.* [22] developed two pilot assignment methods for minimizing the impact of pilot contamination. Time-shifted pilot assignment (TSPA) method were utilized, splitting the cellular network into different groups in which users belonging to the same group were transmitted to uplink-pilots, whereas others received downlink data. Also, the heuristic weighted graph coloring-based pilot assignment (WGC-PA) method was utilized for minimizing intra-group interference. The method showed that two consecutive cells belong to the same group in the time-shifted method. Hua and Chang in [23] devised a “cocktail” method for com-

binning two classic countermeasures, namely time-shifted pilots (TSP) and fractional pilot reuse methods for addressing the issue of pilot contamination in a full-duplex massive MIMO model. The TSP method was adapted for duplicating inner cell regions, such that the count of gathered users is twofold. Also, the base station must function in the full-duplex mode and the symbols of the pilot are employed for attaining the user diversity gain. The outcomes of the simulation verified the accuracy and enhanced the overall uplink and downlink rates. Salh *et al.* [24] developed a channel estimation method which relied on comprehensive knowledge of huge-scale expansion by adapting the orthogonal pilot sequence for eliminating pilot contamination among edge users with minimized channel quality losses, using huge-scale fading. Channel quality available to users was improved. The performance of the methods was analyzed and it was shown that performance losses were minimized, with better channel approximation leading to elevated data rates. However, the method was unable to deal with complex signal processing methods. Wei *et al.* [25] devised a method using time-shifted pilots based on joint position-power optimization for massive MIMO. Here, an effective user scheduling strategy was applied for optimizing pilot power, using the average power constraint. Furthermore, the power and the positions of pilots were optimized iteratively. The analysis performed revealed that the method offered substantial coordination gains in terms of performance, based on spectral efficiency.

2.1. Research Challenges

The research challenges confronted by the conventional time-shifted pilot-based scheduling schemes are:

- in [22], the time-shifted pilot assignment (TSPA) method is devised for reducing the impacts of pilot contamination. However, the TSPA method suffered from intra-group interference that occurred between the cells of similar groups, and offered no approach to grouping cells and to tackling pilot-to-downlink data interference. Moreover, there is a lack of perfect synchronization between cells;
- in massive MIMO-TDD models, the BS obtains CSI using uplink pilots sent by the user with channel reciprocity. However, owing to lower coherence time rates, the reuse of pilots from user terminals in a different cell is foreseeable, which results in pilot contamination issues. Thus, the evaluation of CSI is correlated and produced imprecise beamforming signals [25];
- in massive MIMO-TDD models, the BS obtains CSI using the uplink pilots sent by the user with channel reciprocity. However, owing to less coherence time, the reuse of pilots from user terminals in a different cell is foreseeable, which results in a pilot

contamination issue. Thus, the evaluation of CSI is correlated and produced imprecise beamforming signals [25];

- the major issue in massive MIMO is related to pilot contamination effects impacting channel estimation. Pilot contamination leads to inaccurate channel estimation in 5G downlink transmissions. Besides, various issues need to be faced while executing a massive MIMO model that included noise contamination affecting spectral efficiency.

3. System Model

Massive MIMO is a fundamental element of upcoming applications of 5G networks. MIMO is a wireless network that permits the transmission and reception of large amounts of data signals at the same time and via the same radio channels. In order to satisfy the high data rates and to ensure improved service quality, massive MIMO is favored as it relies on huge antennas at access points. Figure 1 illustrates a massive MIMO model with pilot contamination in the uplink transmission. It comprises antenna arrays, base stations (BS), and different user terminals. The BS sends data packets from various antennas to different terminals at a similar frequency. Massive MIMO [21] is adopted to offer enhanced coverage, high energy efficiency and high data rates. However, various performance restricting issues need to be confronted in this approach in connection with channel estimation and pilot design.

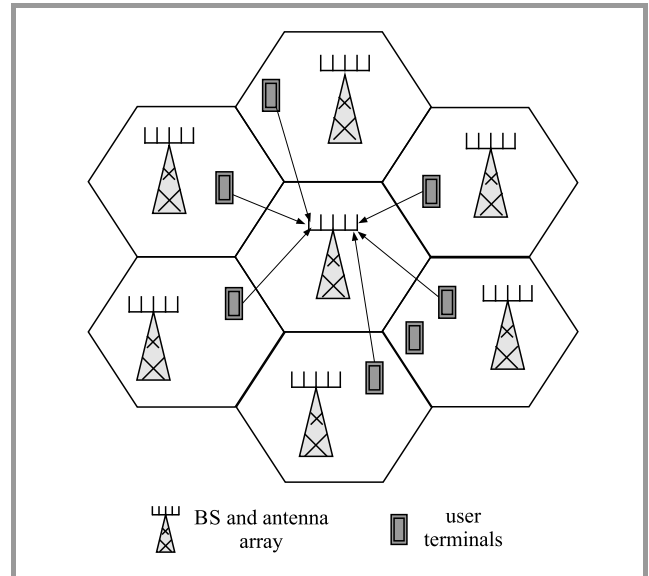


Fig. 1. Model of a massive MIMO system with uplink transmission pilot contamination.

Thus, an effective method is devised for eliminating pilot contamination in order to solve pilot-related problems. It stems from Fig. 1 that each hexagonal cell utilizes the central BS associated with A antennas for D users ($D \gg A$).

Thus, the channel propagation matrix for every user in the cell communicating with the cell site is:

$$b_{def} = a_{def} \sqrt{C_{def}}, \quad (1)$$

where a_{def} indicates the small scale fading vector, $C_{def} = d(d_{def1}, c_{def2}, c_{def3}, \dots, c_{defK})$ is a $K \times K$ order diagonal matrix that describes the large-scale fading factor of each user to the BS of e -th cell. The large-scale fading factor of f -th user of e -th cell to BS of d -th cell is:

$$c_{defK} = \left(\frac{M_{defK}}{k_{defK}} \right)^\eta, \quad (2)$$

where M_{defK} describes shadow fading, k_{defK} is the distance evaluated between f -th user in e -th cell and BS in d -th cell, N represents the radius of the cell, and η is the path loss factor.

The channel vector is given by a constant at coherence time and varies autonomously in varying intervals. Here, the BS produces a downlink channel based on the uplink pilot sequence in coherence time. When data is sent, all users send uplink data to the BS and get downlink at d -th cell as:

$$s_d^u = \sqrt{K_u} \sum_{e=1}^T \sum_{f=1}^D b_{def} x_{ef}^u + v_d^u, \quad (3)$$

where symbol sent from f -th user in e -th cell is x_{ef}^u with $\mathcal{E}\{|x_{ef}^u|^2\} = 1$, and uplink is K_u .

Let $\theta_d = (\theta_1, \theta_2, \dots, \theta_D)^D$ be the pilot sequence matrix of all users in the e -th cell, which satisfies $\theta_d \theta_d^R = X_D$. The pilot sequence matrix received in BS at d -th cell is:

$$s_d^I = \sqrt{N_t} \left(\sum_{d=1}^T b_{def} \theta_D \right) + v_d^u, \quad (4)$$

where N_t is the pilot signal transmission power and v_d^u is the additive white Gaussian noise matrix. For this scenario, the BS of d -th cell evaluates channel matrix R_{dd} with convolutional channel estimation and is:

$$\hat{b}_{def}^{\wedge} = \frac{1}{\sqrt{N_t}} s_d^I \theta_d^R = b_{def} + \sum_{e \neq d} b_{def} + \frac{1}{\sqrt{N_t}} v_d^u \theta_d^R. \quad (5)$$

After the user sends the pilot, every user transmits the data signal to the BS and utilizes the same time frequency resource. The linear blend of channel b_{def} for $1 \leq e \leq T$ is expressed as d_{ddf} , which contains channel users in other cells associated with pilot sequences and is termed pilot contamination. Then, the determined symbol generated from f -th user in d -th cell using a low complexity matched filter detector, offers computed channel matrix $\hat{b}_{def}^{\wedge}$ given as:

$$\hat{x}_{df}^u = b_{ddf}^R s_d^u, \quad (6)$$

$$\hat{x}_{df}^u = \left(\sum_{e=1}^T b_{def}^F + w_{df}^R \right) \left(\sqrt{K_u} \sum_{e=1}^T \sum_{f=1}^D d_{def} x_{ef}^u + v_d^u \right), \quad (7)$$

$$\hat{x}_{df}^u = \sum_{e=1}^T b_{def}^R \cdot \sqrt{K_u} \sum_{e=1}^T \sum_{f=1}^D b_{def} x_{ef}^u + \rho_{d,f}^u, \quad (8)$$

$$\hat{x}_{df}^u = \sqrt{K_u} \left(b_{def}^R b_{ddf} x_{df}^u + \sum_{e \neq d} b_{def}^R b_{def} x_{ef}^u \right) + \rho_{d,f}^u, \quad (9)$$

$$\hat{x}_{df}^u = A \sqrt{K_u} \left(\alpha_{ddf} x_{df}^u + \sum_{e \neq d} \alpha_{def} x_{ef}^u \right) + \rho_{d,f}^u, \quad (10)$$

where w_{dc} indicates f -th column of $\frac{v_d^I \theta_d^R}{\sqrt{s_j}}$ and $\rho_{d,f}^u$ represents orthogonality between channels of many users. When the BS antenna count reaches to infinity, the detected symbol is:

$$x_{df}^u \approx A \sqrt{K_u} \left(\alpha_{ddf} x_{df}^u + \sum_{e \neq m} \alpha_{def} x_{ef}^u \right). \quad (11)$$

4. Scheduling Using Adaptive ESMO Algorithm

The proposed scheduling algorithm utilizes time-shifted pilots. At first, the grouping of users is performed based on the user grouping algorithm which assists in grouping users (center user and edge user) based on various levels of contaminating pilots. The grouping of users is performed using sparse FCM, based on such parameters as large-scale fading factor, SINR, and user distance. Finally, pilot scheduling is performed using adaptive ESMO, i.e. a method that combines adaptive tuning parameters with ESMO. Figure 2 shows a schematic view of the time-shifted pilot-based scheduling process using the proposed adaptive ESMO algorithm.

4.1. Grouping of Users with Sparse FCM

Users are divided into different groups in order to minimize the complexity of processing by serving each group individually. In a specific time slot, a subset of users is scheduled within each group for transmission. Additionally, the precise SINR of each user is essential for user grouping. However, in conventional user grouping methods, channel covariance matrices are adopted for grouping users, which causes inter-group interference. Here, user computing parameters and sparse FCM are employed for facilitating the grouping process. The parameters include user distance, SINR, and the large-scale fading factor [21]. These parameters are essential for designing the user grouping method in order to improve system capacity.

The uplink SINR of v -th user in b -th cell is:

$$\text{SINR}_{df}^u = \frac{|b_{ddf}^R b_{ddf}|^2}{\sum_{e \neq d} |b_{def}^R b_{def}|^2 + \frac{|\rho_{d,d}^u|^2}{K_u}}, \quad (12)$$

With large numbers of BS, the uplink SINR is expressed as:

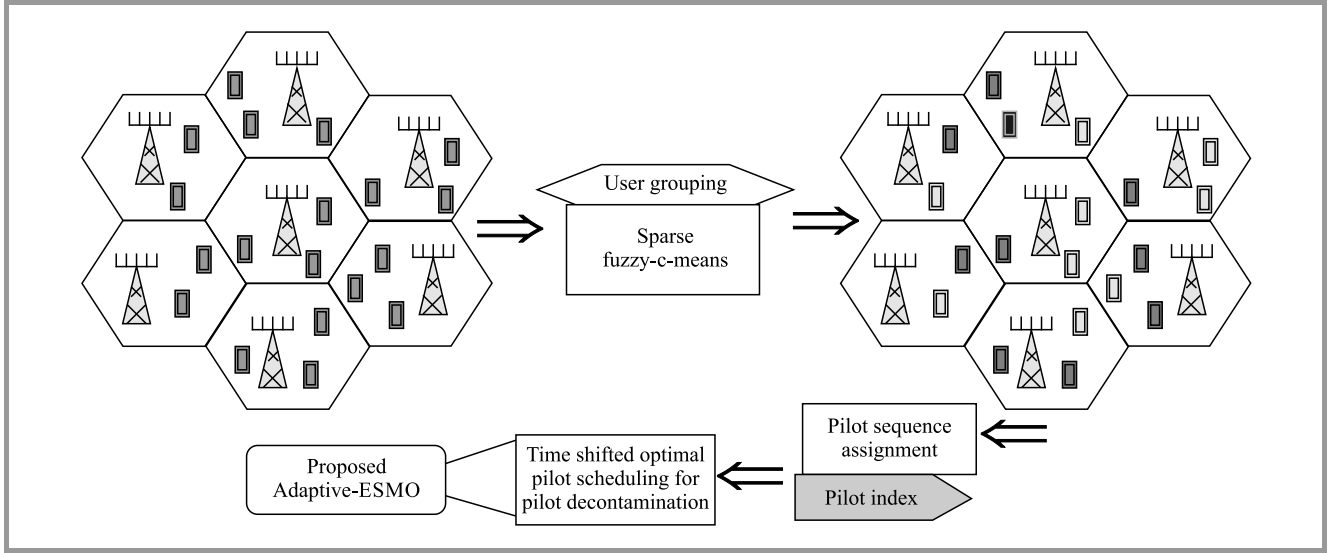


Fig. 2. Schematic view of the time-shifted pilot-based scheduling using proposed adaptive ESMO.

$$SINR_{df}^u = \frac{\alpha_{ddf}^2}{\sum_{e \neq d} \alpha_{def}^2}. \quad (13)$$

Hence, the SINR of the user is expressed as G_1 .

The large scaling factor of f -th user of e -th cell to BS of d -th cell is obtained in Eq. (2). This factor is expressed as G_2 . User distance is computed between the location of f -th user and the center of the cell:

$$\partial = \sqrt{\sum_{f=1}^L \left(\omega_f^{loc}, z^{cell} \right)}, \quad (14)$$

where ∂ indicates the user distance, ω_f^{loc} indicates the location of f -th user, z^{cell} represents the center of the cell, and G_3 indicates user distance. Hence, user grouping-related parameters are formulated as:

$$T = G_1, G_2, G_3. \quad (15)$$

User grouping is performed with sparse FCM – a method that relies on such parameters as large-scale fading factor, SINR, and user distance. The goal of sparse FCM is to group users into center user and edge user categories. In sparse FCM [27], dimensional reduction is essential. Here, the advantage consists in grouping users without any delay, and in reducing the level of complexity caused by the optimization method. The steps of sparse FCM are:

- **Initialization.** The foremost step is to initialize the feature weights, which are:

$$\varpi = \varpi_1^n = \varpi_2^n = \dots = \varpi_z^n = \frac{1}{\sqrt{z}}, \quad (16)$$

where z indicates features and n -th attribute weights based on z -th feature indicated as ω_z^n ;

- **Update the partition matrix.** Set cluster centers and attribute weights as q and ω . The constraint is $\vartheta(W)$, which must be reduced to:

$$\omega_{ic} = \begin{cases} \frac{1}{I_c} & \text{if } r_{ic} \text{ and } I_c = \text{cardinality}\{e : r_{ic} = 0\} \\ 0 & \text{if } r_{ic} \neq 0 \text{ but } r_{iy} = 0 \text{ for some } y, y \neq l \\ \frac{1}{\sum_{y=1}^o \left(\frac{r_{ic}}{r_{yc}} \right)^{\left(\frac{1}{\gamma-1} \right)}} & \text{otherwise} \end{cases}, \quad (17)$$

where $r_{ic} = \sum_{c=1}^o \varpi_e (v_{ei} - k_{ci})^2$ and $\text{cardinality}(U)$ represent cardinality of set U .

The distance measure is r_{ic} which indicates the distance between i -th node and c -th cluster center;

- **Determine cluster centers.** For evaluating cluster centers, it is important to set attribute weight ω and data matrix Z , and estimation of cluster center is:

$$y_{ce} = \begin{cases} 0 & \text{if } \varpi_e = 0 \\ \frac{\sum_{i=1}^n \omega_{ic}^\gamma v_{ei}}{\sum_{i=1}^n \omega_{ic}^\gamma} & \text{if } \varpi_e \neq 0 \end{cases}, \quad (18)$$

where $c = 1, \dots, v$ and $e = 1, \dots, d$. The c index indicates the cluster center and e represents data attribute. The evaluation of the cluster center is done based on attribute weight ϖ_e . The response of e -th

feature to the objective function is indicated as $\bar{\omega}_e$ and constant Y symbolizes dissimilarity measure;

- **Derive the class.** The class is determined using specified clusters $y = \{y_1, y_2, \dots, y_c, \dots, y_o\}$ having a membership ω . Class Y_e is detected with the use of an objective function:

$$\max_{\bar{\omega}} \sum_{e=1}^z \bar{\omega}_e \cdot Y_e, \quad (19)$$

such that $\|\bar{\omega}\|_2^2 \leq 1$, $\|\bar{\omega}\|_{\kappa}^{\kappa} \leq \lambda$ produces $\bar{\omega}^*$, where λ indicates tuning parameter ($0 \leq \kappa \leq 1$) and $\|\bar{\omega}\|_{\kappa}^{\kappa} = \sum_{e=1}^d |\bar{\omega}_e|^{\kappa}$;

- **Terminate.** The iteration is repeated until the maximum count is reached. The stopping criterion is:

$$\frac{\sum_{e=1}^z |\bar{\omega}_e^* - \bar{\omega}_e^o|}{\sum_{e=1}^z |\bar{\omega}_e^o|} < 10^{-4}. \quad (20)$$

Hence, the count of users in the individual cell is chosen vibrantly. The proposed user grouping approach mitigates inapt grouping of users and facilitates effective grouping even under the worst channel conditions.

Based on the clustering result, the distance attribute of the centroid having the minimum value is considered. The corresponding user belongs to center users (group 1), and the remaining user belongs to edge users (group 2) or nearby edge users (group 3).

4.2. Time-shifted Optimal Pilot Scheduling

Time-shifted pilot slot allocation using the adaptive ESMO method is devised for alleviating impacts of contaminating pilots. The method relies on four time-based scheduling phases, including pilot sequence transmission, processing and estimation of the channel at BS, uplink data transmission, and downlink data transmission. This method assigns pilots to cells at different time-shifted slots. Non-overlapping time slots are assigned to the transmission of pilots by shifting pilot locations. Thus, the use of time-shifting is controlled effectively, without any overlaps, and pilot contamination is removed entirely. A brief illustration of the proposed adaptive ESMO used for time-shifted pilot slot allocation is given below.

The goal of solution encoding is to produce a solution vector using the proposed optimization technique. The solution consists of pilots that match the proposed adaptive ESMO method (Fig. 3). The count of pilots is chosen based on a matrix, wherein the count of users is indicated in a row-wise manner and the count of cells is indicated in a column-wise manner. Let us assume that the total number users is U , the count of groups is G and the count of cells is A . Then, the dimension of the solution is indicated as $A \times G$. Based on the matrix, a solution vector is described

that indicates the count of pilots, and its size is represented as $(A \times G)$.

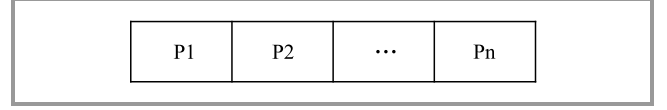


Fig. 3. Solution encoding to pilot assignment.

Let us assume that $U = 6$, $G = 3$, and $A = 2$. Let the number of pilots be 2, meaning that pilot length becomes 2. The dimension of the solution is $2(A)$. Figure 4 shows an example of the pilot assignment process.

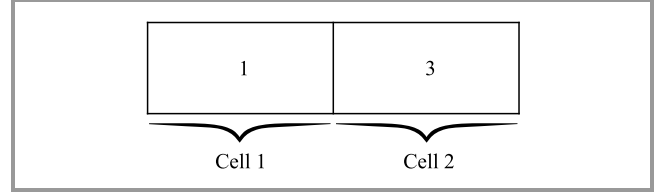


Fig. 4. Example of pilot assignment process.

Pilot p_1, p_2 is assigned to cell 1 with the index of 1 and 3. The mapping of the above solution in the concept model using a time-shifted pilot sequence is illustrated in Fig. 5. Let us assume that $P, D, U = \{2, 4, 4\}$, where P indicates the total number of pilots, D denotes the downlink scenario and U is an uplink scenario. Then, the solution matrix is represented as:

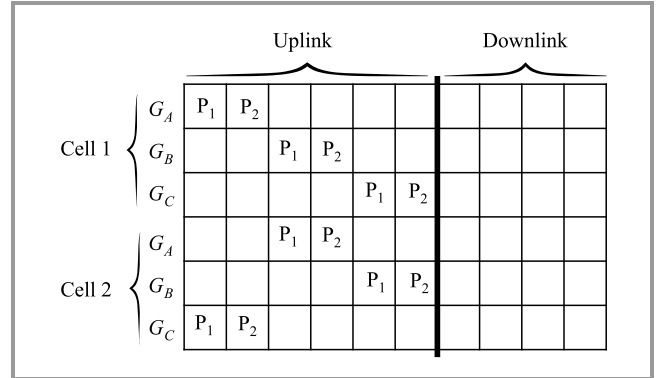


Fig. 5. Example of the time-shifted pilot sequence.

The level of fitness helps reveal solution quality worked out with the use of the achievable sum rate. It is modeled as a maximization function and is derived below.

The utility function describes the achievable uplink rate, wherein the pilot sequence θ_v is assigned to the E_f -th user:

$$\Theta(E_f, \theta_f) = P_{E_f}^u(E_f, \theta_f), \quad (21)$$

where pilot sequence θ_f allocated to user E_f is expressed as $(E_f, \theta_f) \{W_v : v = 1, \dots, D!\}$ represents all pilot assignment possibilities, and $W_v = [W_v^1, W_v^2, \dots, W_v^D]$ indicates v -th assignment. Furthermore, the goal of massive MIMO

is to enhance the achievable target cell sum rate, which is the main optimization issue and is given by:

$$Z : \max_{\{W_v\}} \left\{ \sum_{D=1}^E P_{E_f}^u(E_f, \theta_{W_v^u}) \right\}. \quad (22)$$

The enhanced achievable uplink rate is:

$$P(E, \theta) = (1 - \mu_0) \left\{ \log_2 [1 + \text{SINR}(E, \theta)] \right\} + \mu_0 \times \sum_{a=1}^n B_a, \quad (23)$$

where $\text{SINR}(E, \theta)$ indicates SINR value θ at user E and he transmit power of pilot is expressed as B_a . The total pilot is indicated as a . ESMO provides a solution to the optimization issue by offering optimum pilot assignments.

4.3. Steps of Proposed Adaptive ESMO

Time-shifted optimal pilot scheduling is performed using the proposed adaptive ESMO algorithm which is a combination of ESMO and adaptive tuning parameters. Here, ESMO is determined by integrating EHO from [31] and SMO [39]. The issues of EHO are addressed with SMO that tends to offer a better convergence rate by generating a globally optimized solution. Such a method is simple and may effectively manage several local optima, significant nonlinearity, ruggedness, and interdependence. Additionally, it paves the way to increasing robustness of the optimization technique.

In the self-adaptive method, selection of the learning strategy and control parameters is not required beforehand, and parameter settings self-adapt based on the learning experience. Thus, the incorporation of adaptive concepts into the ESMO algorithm improves the overall performance. The steps of the proposed adaptive ESMO algorithm are as follows:

Initialization. The initialization of the solutions in the search space is:

$$S = \{S_1, S_2, \dots, S_\kappa, \dots, S_\ell\}, \quad (24)$$

where S_κ indicates κ -th solution, and ℓ is the total number of solutions.

Evaluation of fitness function. The optimum solution is detected using the fitness function. The solution with maximum fitness is chosen as optimal.

Determination of update solution. After computing their fitness, the solutions update ESMO, which enhances the performance of the algorithm as far as discovering optimal pilots is concerned:

$$S_{e,f}(y+1) \left[\frac{[\mu - 1] [S^*(y) - S_{e,f}(y)]}{\mu \times [S^*(y) - S_{e,f}(y)]} \right] = S_{e,f}(y) - \frac{S_{e,f}(y) [S^*(y) - S_{e,f}(y)]}{\mu \times [S^*(y) - S_{e,f}(y)]} + T(-1, 1) \times [S_{o,f}(y) - S_{e,f}(y)], \quad (25)$$

where $S_{e,f}(y)$ indicates the f dimension of e -th SM, $S_{o,f}(y)$ refers to the f dimension of o -th SM, the random number is denoted as $T(-1, 1)$, $S^*(y)$ signifies the best solution, and μ denotes the scale factor ranging from 0 to 1. The update is performed using the ESMO algorithm by making the constants self-adaptive in ESMO. To incorporate the adaptive concept, μ is made self-adaptive, such as:

$$\mu = \frac{1}{4} \left[\frac{\text{Absolute rate}}{M} + \frac{t}{T} + \left(\frac{\text{Number of index changed in the count at } t\text{-th iteration}}{d} \right) z \right], \quad (26)$$

where T indicates total time, M is constant, t is iteration index, d is a diameter. The solution update obtained in Eq. (26) is used for time-shifted optimal pilot scheduling.

Determination of best solution. The possible solutions are ranked based on the fitness function and the one that ranks the highest is the best solution.

Terminate. The best solutions are obtained iteratively until the maximum number of iterations are attained.

5. Results and Discussion

The proposed adaptive ESMO approach was analyzed using Matlab and was compared with existing methods based on achievable rate, SINR, and throughput by varying the number of users, the number of antennas and log-normal shadowing fading. The evaluated parameters include achievable rate, SINR, and throughput.

The achievable sum is used to optimize pilot information acquired from the channel dynamics and is:

$$A_r = B \times \log_2 \left(1 + \frac{P_s}{N + \sum_{i \in L} I^i} \right), \quad (27)$$

where B indicates bandwidth, P_s refers to received signal power, N symbolizes background noise, and I^i indicates receiver interference.

SINR is the proportion of signal power and the sum of interference power from additional interfering signals with background noise:

$$\text{SINR} = \frac{\sigma}{\kappa + \chi}, \quad (28)$$

where σ indicate the incoming signal, κ represents interference power, and χ is a constant.

The throughput of the network is the rate of successful exchange of data through a communication channel within a specific period.

5.1. Experimental Results

Figure 6 shows the experimental results of the proposed adaptive ESMO with 10 and 15 users. The red triangle indicates BS, whereas the green circle represents mobile devices. Communication between users and BS involves the sending of pilots to evaluate channels. The user sends pilots at the same time, which results in pilot contamination. Thus, pilot contamination is managed by the adaptive ESMO model using time-shifted pilot based scheduling.

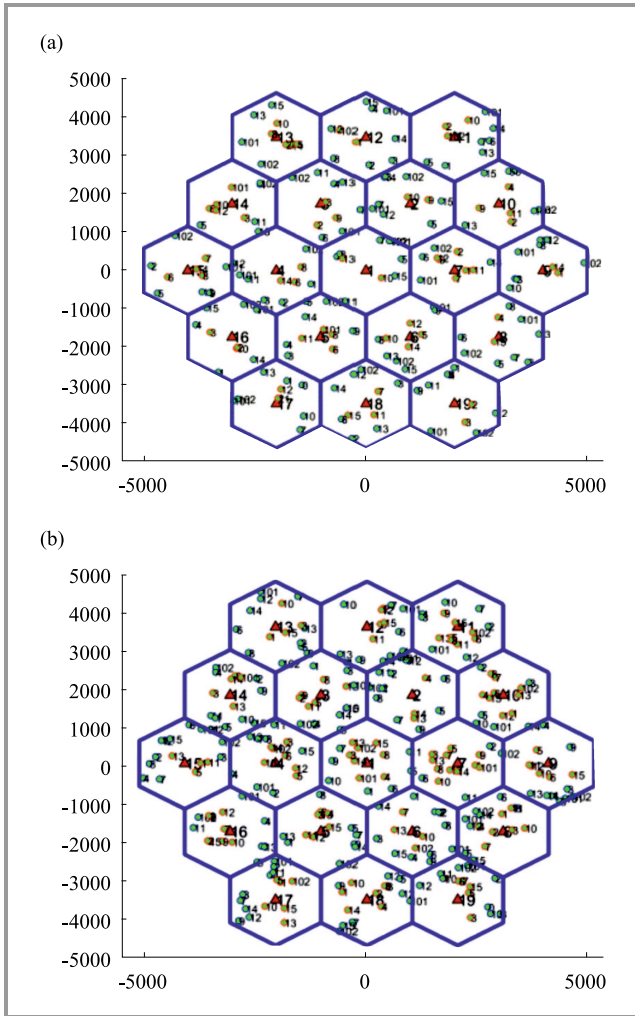


Fig. 6. Experimental results of the proposed adaptive ESMO using: (a) 10 users, (b) 15 users. (For color pictures see the digital version of the paper).

The remaining techniques used here for the purpose of the comparative analysis include: degradation-based scheduling from [26], MRT-ZF presented in [30], soft pilot reuse

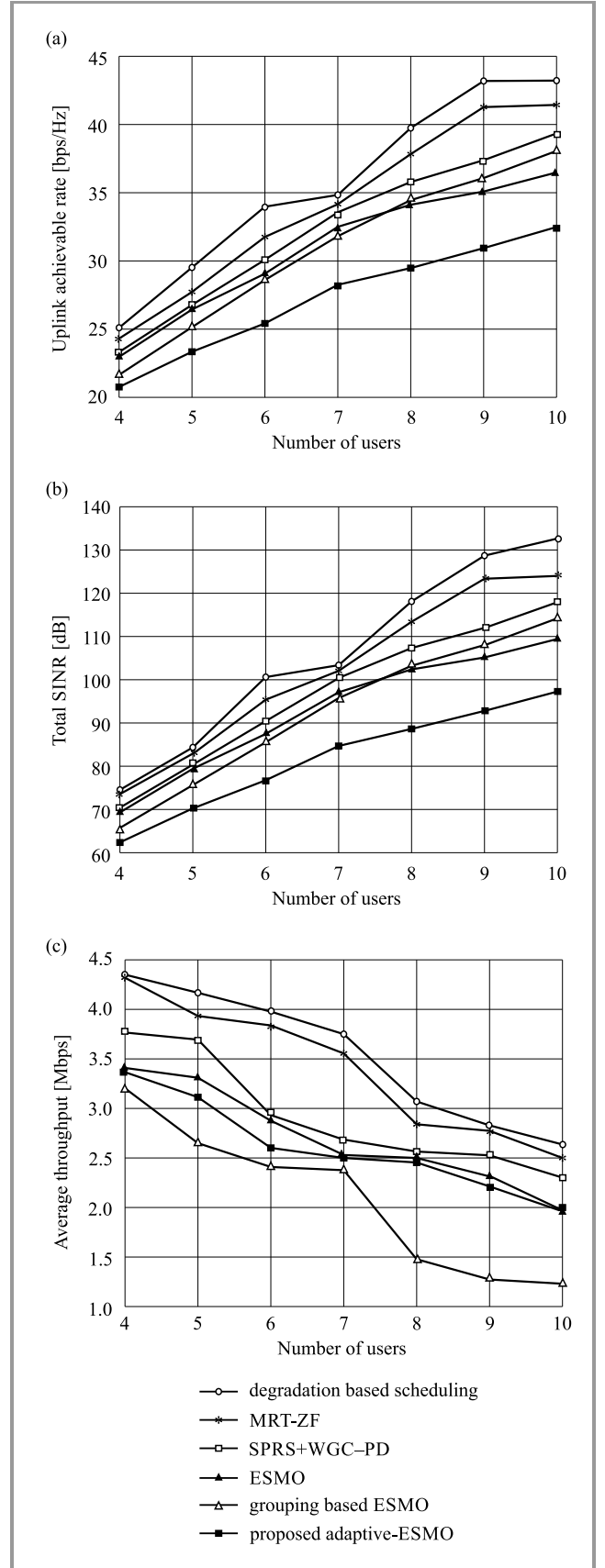


Fig. 7. Comparative analysis based on altering the number of users: (a) uplink achievable rate, (b) total SINR, and (c) average throughput.

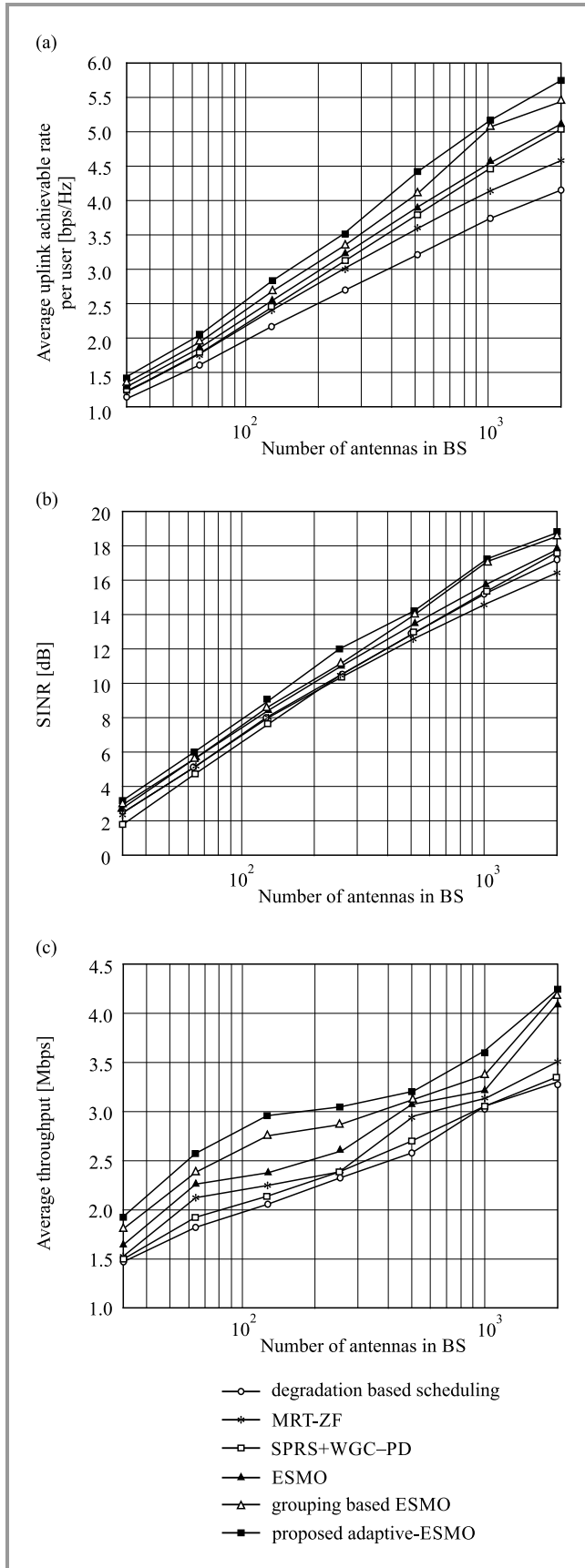


Fig. 8. Comparative analysis based on varying the number of the antennas in BS: (a) uplink achievable rate, (b) total SINR, and (c) average throughput.

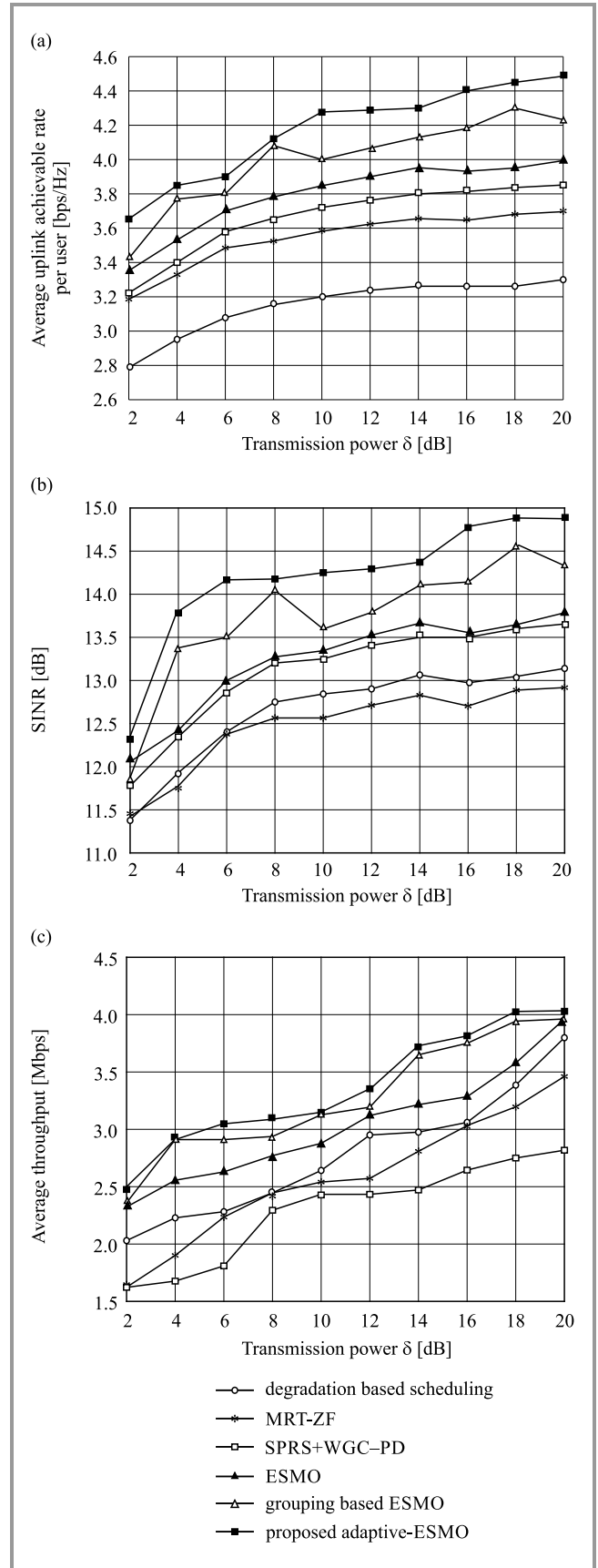


Fig. 9. Comparative analysis based on altering transmission power: (a) uplink achievable rate, (b) total SINR, and (c) average throughput.

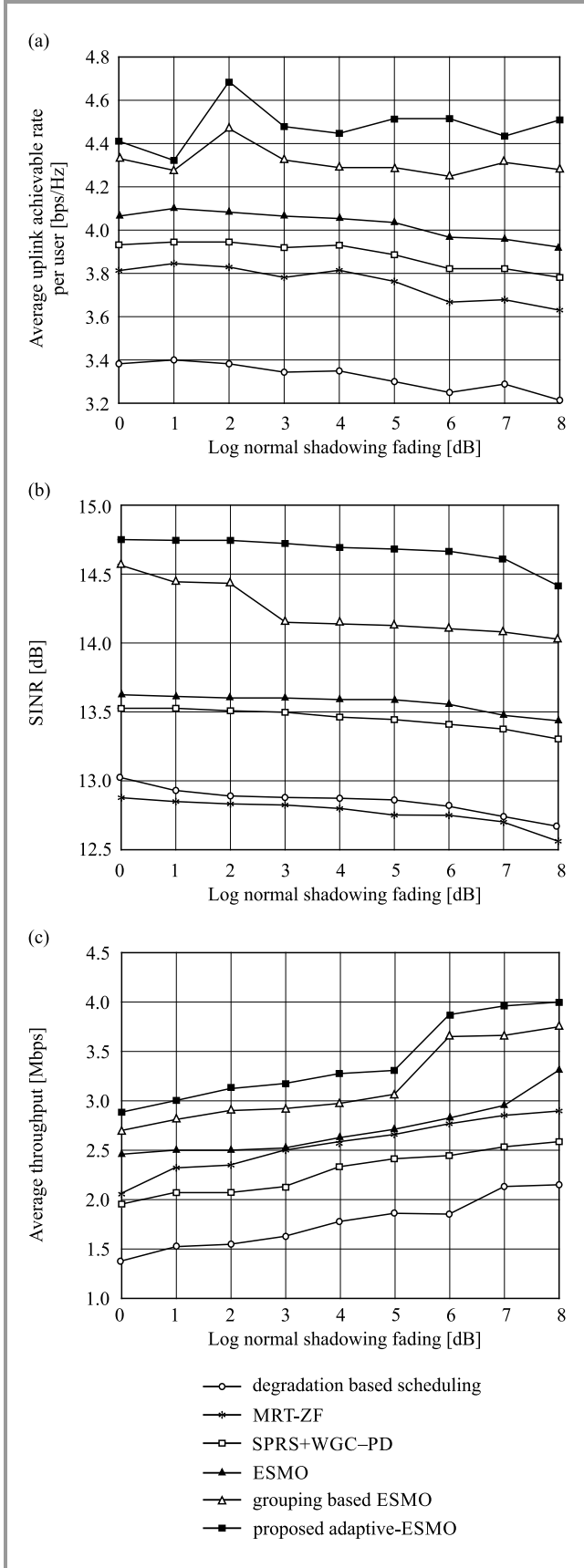


Fig. 10. Comparative analysis based on altering log-normal shadowing fading: (a) uplink achievable rate, (b) total SINR, and (c) average throughput.

scheme + weighted coloring graph + pilot decontamination (SPRS+WGC-PD) from [31], ESMO, grouping-based ESMO and proposed adaptive ESMO.

Figure 7 illustrates the analysis of methods by altering the count of users considering the uplink achievable rate, total SINR, and average throughput.

Figure 8 shows an analysis of the different methods based on altering the number of antennas in BS, with the same criteria adopted.

Figure 9 presents an analysis of the methods based on altering the transmission power, while Fig. 10 shows alterations of log-normal shadowing fading, performed with the use of the same parameters as used in Figs. 7–8.

5.2. Discussion Concerning Achieved Results

Table 1 presents an analysis of the different methods considering the achievable uplink rate, total SINR, and average throughput. Considering the number of users, the maximum achievable rate of 43.084 bps/Hz is computed for adaptive ESMO. In contrast, the existing methods showed lower achievable rates. The maximum SINR of adaptive ESMO is 132.882 dB, whereas SINR of the other methods equals: degradation-based scheduling – 97.438 dB, MRT-ZF – 109.633 dB, SPRS+WGC-PD – 114.386 dB, ESMO – 118.176 dB, and grouping-based ESMO – 124.318 dB. The maximum average throughput computed for adaptive ESMO is 2.633 Mbps. Considering the number of antennas in BS, the proposed adaptive ESMO reached the maximum achievable rate of 5.758 bps/Hz, total SINR of 18.843 dB, and average throughput of 4.229 Mbps. For transmission power, adaptive ESMO reached the maximum achievable rate of 4.490 bps/Hz, total SINR of 14.874 dB, and average throughput of 4.022 Mbps. In the case of log-normal shadowing fading, adaptive ESMO reached the maximum achievable rate of 4.513 bps/Hz, total SINR of 14.409 dB, and average throughput of 4.017 Mbps.

6. Conclusion

In order to mitigate the adverse impact of pilot contamination, time-shifted pilot scheduling is performed using adaptive ESMO, a method incorporating adaptive tuning parameters into the ESMO algorithm. Here, the grouping of users is performed based on attributes including large-scale fading factor, SINR, and user distance. The sparse-FCM algorithm is used as well, helping categorize users as center users or edge users. The grouping of users is performed in such a way that the method effectively alleviates inappropriate grouping of users, and facilitates effective grouping, even under the worst channel conditions. Finally, pilot sequences are subjected to time-shifted pilot scheduling using the proposed adaptive ESMO algorithm in order to eliminate pilot contamination. The effectiveness of the proposed method was evaluated using specific parameters and revealed superior performance with

Table 1
Comparative data retrieved by the simulations

Parameter	Metric	Degradation-based scheduling	MRT-ZF	SPRS+WGC-PD	ESMO	Grouping-based ESMO	Proposed adaptive ESMO
Number of users	Achievable rate [bps/Hz]	32.368	36.419	37.998	39.257	41.298	43.084
	Total SINR [dB]	97.438	109.633	114.386	118.176	124.318	132.882
	Average throughput [Mbps]	1.945	1.221	1.951	2.280	2.495	2.633
Number of the antennas in BS	Achievable rate [bps/Hz]	4.157	4.582	5.055	5.120	5.448	5.758
	Total SINR [dB]	17.353	16.552	17.730	17.877	18.679	18.843
	Average throughput [Mbps]	3.294	3.515	3.347	4.076	4.194	4.229
Transmission power	Achievable rate [bps/Hz]	3.301	3.689	3.850	3.988	4.227	4.490
	Total SINR [dB]	13.127	12.915	13.658	13.791	14.315	14.874
	Average throughput [Mbps]	3.804	3.463	2.820	3.949	3.962	4.022
Log-normal shadowing fading	Achievable rate [bps/Hz]	3.227	3.642	3.785	3.927	4.286	4.513
	Total SINR [dB]	12.676	12.560	13.311	13.437	14.035	14.409

the highest achievable uplink rate of 43.084 bps/Hz, the highest SINR of 132.882 dB, and maximum throughput of 2.633 Mbps, respectively.

References

- [1] I. A. Khan, "Robust signal detection scheme for 5G massive multiuser MIMO systems", *IEEE Trans. Veh. on Technol.*, vol. 67, pp. 9597–9604, 2018 (DOI: 10.1109/TVT.2018.2858922).
- [2] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays", *IEEE Sig. Process.*, vol. 30, pp. 40–60, 2013 (DOI: 10.1109/MSP.2011.2178495).
- [3] S. Biswas, J. Xue, F. A. Khan, and T. Ratnarajah, "Performance analysis of correlated massive MIMO systems with spatially distributed users", *IEEE Syst. J.*, vol. 12, pp. 1850–1861, 2016 (DOI: 10.1109/JSYST.2016.2594155).
- [4] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas", *IEEE Trans. on Wirel. Commun.*, vol. 9, no. 11, pp. 3590–3600, 2010 (DOI: 10.1109/TWC.2010.092810.091092).
- [5] X. Zhu, L. Dai, and Z. Wang, "Graph coloring based pilot allocation to mitigate pilot contamination for multi-cell massive MIMO systems", *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1842–1845, 2015 (DOI: 10.1109/LCOMM.2015.2471304).
- [6] W. A. Mahyiddin, P. A. Martin, and P. J. Smith, "Pilot contamination reduction using time-shifted pilots in finite massive MIMO systems", in *Proc. 2014 IEEE 80th Vehicular Technol. Conf. VTC2014-Fall 2014*, Vancouver, BC, Canada, 2014 (DOI: 10.1109/VTCFall.2014.6966130).
- [7] H. Yin, D. Gesbert, M. C. Filippou, and Y. Liu, "Decontaminating pilots in massive MIMO systems", in *Proc. of IEEE Int. Conf. on Commun. ICC 2013*, Budapest, Hungary, 2013, pp. 3170–3175 (DOI: 10.1109/ICC.2013.6655031).
- [8] R. Muller, L. Cottatellucci, and M. Vehkaperä, "Blind pilot decontamination", *IEEE J. of Sel. Topics in Sig. Process.*, vol. 8, no. 5, pp. 773–786, 2014 (DOI: 10.1109/JSTSP.2014.2310053).
- [9] R. Mochaourab, E. Björnson, and M. Bengtsson, "Pilot clustering in asymmetric massive MIMO networks", in *Proc. of the IEEE 16th Int. Worksh. on Sig. Process. Adv. in Wirel. Commun. SPAWC 2015*, Stockholm, Sweden, 2015 (DOI: 10.1109/SPAWC.2015.7227034).
- [10] A. Ashikhmin and T. L. Marzetta, "Pilot contamination precoding in multi-cell large-scale antenna systems", in *Proc. of the IEEE Int. Symp. on Inform. Theory Proceedings ISIT 2012*, Cambridge, MA, USA, 2012, vol. 1, pp. 1137–1141, 2012 (DOI: 10.1109/ISIT.2012.6283031).
- [11] L. Liangbin, A. Ashikhmin, and T. L. Marzetta, "Pilot contamination precoding for interference reduction in large-scale antenna systems", in *Proc. of the 51th Ann. Allerton Conf. on Commun., Control, and Comput. Allerton CCC 2013*, Monticello, IL, USA, 2013, vol. 2, no. 2, pp. 226–232 (DOI: 10.1109/Allerton.2013.6736528).
- [12] T. M. Nguyen and L. B. Le, "Joint pilot assignment and resource allocation in multicell massive MIMO network: Throughput and energy efficiency maximization", in *Proc. of the IEEE Wirel. Commun. and Network. Conf. WCNC 2015*, New Orleans, LA, USA, 2015, vol. 9, pp. 393–398 (DOI: 10.1109/WCNC.2015.7127502).
- [13] H. V. Cheng, E. Björnson, and E. G. Larsson, "Optimal pilot and payload power control in single-cell massive MIMO systems", *IEEE Trans. on Sig. Process.*, vol. 65, no. 9, pp. 2363–2378, 2017 (DOI: 10.1109/TSP.2016.2641381).
- [14] H. T. Dao and S. Kim, "Pilot power allocation for maximizing the sum rate in massive MIMO systems", *IET Commun.*, vol. 12, no. 11, pp. 1367–1372, 2018 (DOI: 10.1049/iet-com.2017.1407).
- [15] S. Jin *et al.*, "On massive MIMO zero-forcing transceiver using time-shifted pilots", *IEEE Trans. on Veh. Technol.*, vol. 65, no. 1, pp. 59–74, 2016 (DOI: 10.1109/TVT.2015.2391192).

- [16] X. Xiong, B. Jiang, X. Gao, and X. You, "QoS-guaranteed user scheduling and pilot assignment for large-scale MIMO-OFDM systems", *IEEE Trans. on Veh. Technol.*, vol. 65, no. 8, pp. 6275–6289, 2015 (DOI: 10.1109/TVT.2015.2477683).
- [17] X. Dai, "Optimal training design for linearly time-varying MIMO/OFDM channels modeled by a complex exponential basis expansion", *IET Commun.*, vol. 1, no. 5, pp. 945–953, 2007 (DOI: 10.1049/iet-com:20045301).
- [18] B. Akgun, M. Krunz, and O. O. Koyluoglu, "Vulnerabilities of massive MIMO systems to pilot contamination attacks", *IEEE Trans. on Inform. Foren. and Secur.*, vol. 14, no. 5, pp. 1251–1263, 2018 (DOI: 10.1109/TIFS.2018.2876750).
- [19] Y. Wu, C.-K. Wen, W. Chen, S. Jin, R. Schober, and G. Caire, "Data-aided secure massive MIMO transmission under the pilot contamination attack", *IEEE Trans. on Commun.*, vol. 67, no. 7, pp. 4765–4781, 2019 (DOI: 10.1109/TCOMM.2019.2907943).
- [20] J. Fan, W. Li, and Y. Zhang, "Pilot contamination mitigation by fractional pilot reuse with threshold optimization in massive MIMO systems", *Digit. Sig. Process.*, vol. 78, pp. 197–204, 2018 (DOI: 10.1016/j.dsp.2018.02.011).
- [21] O. A. Saraereh, I. Khan, B. M. Lee, and A. Tahat, "Efficient pilot decontamination schemes in 5G massive MIMO systems", *Electronics*, vol. 8, no. 1, pp. 55, 2019 (DOI: 10.3390/electronics8010055).
- [22] I. E. Shaalan, A. A. Khattaby, and A. S. Dessouki, "A new joint TSPA/WGC pilot contamination reduction strategy based on exact graph coloring grouping algorithm", *IEEE Access*, vol. 7, pp. 150552–150564, 2019 (DOI: 10.1109/ACCESS.2019.2947665).
- [23] Y. K. Hua and W. Chang, "Time shifted pilots scheme for full-duplex massive MIMO systems", *IEEE Trans. on Veh. Technol.*, vol. 68, no. 3, pp. 3022–3026, 2019 (DOI: 10.1109/TVT.2019.2893547).
- [24] A. Salh, L. Audah, N. S. M. Shah, and S. A. Hamzah, "Mitigating pilot contamination for channel estimation in multi-cell massive MIMO systems", *Wirel. Pers. Commun.*, vol. 112, pp. 1643–1658, 2020 (DOI: 10.1007/s11277-020-07120-9).
- [25] T. Wei, W. Feng, N. Ge, and J. Lu, "Optimized time-shifted pilots for maritime massive MIMO communication systems", in *Proc. 26th Wirel. and Opt. Commun. Conf. WOCC 2017*, Newark, NJ, USA, 2017 (DOI: 10.1109/WOCC.2017.7929003).
- [26] Y. Wu, T. Liu, M. Cao, L. Li, and W. Xu, "Pilot contamination reduction in massive MIMO systems based on pilot scheduling", *EURASIP J. on Wirel. Commun. and Network.*, 2018, article no. 21, 2018 (DOI: 10.1186/s13638-018-1029-1).
- [27] X. Chang, Q. Wang, Y. Liu, and Y. Wang, "Sparse regularization in fuzzy c-means for high-dimensional data clustering", *IEEE Trans. on Cybernet.*, vol. 47, no. 9, pp. 2616–2627, 2017 (DOI: 10.1109/TCYB.2016.2627686).
- [28] G. G. Wang, S. Deb, and L. D. S. Coelho, "Elephant herding optimization", in *Proc. of 3rd Int. Symp. on Comput. and Business Intell. ISCBI 2015*, Bali, Indonesia, 2015 (DOI: 10.1109/ISCBI.2015.8).
- [29] J. C. Bansal, H. Sharma, S. S. Jadon, and M. Clerc, "Spider monkey optimization algorithm for numerical optimization", *Memetic Comput.*, vol. 6, no. 1, pp. 31–47, 2014 (DOI: 10.1007/s12293-013-0128-0).
- [30] J. Li, D. Wang, P. Zhu, J. Wang, and X. You, "Downlink spectral efficiency of distributed massive MIMO systems with linear beamforming under pilot contamination", *IEEE Trans. on Veh. Technol.*, vol. 67, no. 2, pp. 1130–1145, 2018 (DOI: 10.1109/TVT.2017.2733532).
- [31] W. Yuan, X. Yang, and R. Xu, "A novel pilot decontamination scheme for uplink massive MIMO systems", *Procedia Comp. Sci.*, vol. 131, pp. 72–79, 2018 (DOI: 10.1016/j.procs.2018.04.187).



Ambala Pradeep Kumar received his B.Tech. degree and M.Tech. from Jawaharlal Nehru Technological University, Hyderabad, India, in 2008 and 2012, respectively. Between 2008 and 2016 he worked as an Assistant Professor at SR Engineering College, Warangal, India. He is currently pursuing Ph.D. at the Department of ECE, KU College of Engineering and Technology, Kakatiya University, Warangal, India. He published 6 papers in international journals and attended numerous conferences. His research interests are in the field of digital signal processing and wireless communications.
E-mail: ambalapradeepk@gmail.com
Department of ECE
KU College of Engineering and Technology
Kakatiya University
Warangal, Telangana, India



Tadisetty Srinivasulu is working as Professor and Dean at the Faculty of Engineering and Technology, Kakatiya University. He obtained M.Tech. from IIT Dhanbad, and Ph.D. from Kyushu University, Japan. He completed Proficiency Courses and received diplomas from IISc Bangalore, IIT Kharagpur, CEDTI Gorakhpur, NPA and NRDC New Delhi. The major administrative positions he held include Principal, Director, General Manager, Coordinator and Head of the Department. He published over 157 research papers in international journals and conference papers.
E-mail: drstadisetty@gmail.com
Department of ECE
KU College of Engineering and Technology
Kakatiya University
Warangal, Telangana, India

MIMO Antenna Design and Optimization with Enhanced Bandwidth for Wireless Applications

Mekala Harinath Reddy¹ and D. Sheela²

¹ School of Engineering, University of Leicester, Leicester, United Kingdom

² Department of ECE, Tagore Engineering College, Chennai, India

<https://doi.org/10.26636/jtit.2020.145520>

Abstract—This paper demonstrates a compact MIMO (Multi Input Multi Output) fractal type antenna for ultra-wide band applications. The proposed antenna is manufactured on a low-cost substrate material and the design is analyzed for various iterations in terms of reflection coefficient, gain, and bandwidth. The 50 Ω transmission line feed is used for both fractal patches and a metamaterial structure is used as the ground plane. The proposed design achieved a wide-band frequency response between 5.8 and 15 GHz, with the reflection coefficient of less than -10 dB. Reduced mutual coupling, positive gain and stable radiation patterns were observed throughout the operating band as well. The bandwidth of 9.2 GHz is achieved with the use of a metamaterial structure on the ground plane. The ECC and diversity gain obtained prove the excellent diversity performance of the antenna. The design was simulated using HFSS software and was tested in a lab.

Keywords—HFSS, metamaterial, mutual coupling, permittivity, reflection coefficient

1. Introduction

A microstrip patch antenna is one of the most popular types of printed type aerials. It plays a significant role in today's world of wireless communication [1]. Due to its miniature size, low cost of manufacturing and light weight, such a type is preferred over other antennas in many applications [2]. Microstrip patch antennas offer numerous advantages and many other technologies have been invented for bandwidth enhancement and high gain [3].

A MIMO antenna is one of such technologies, where multiple antennas are used forming a system at the transmitter and the receiver, improving the quality of wireless communication by link multiplication [4]. MIMO is relied upon to explore high performance and low-cost antenna layouts, as well as to support multimedia services with high channel capacities and transmission speeds [5]. The main advantage of this type of antenna is that it is capable of sending and receiving data via multiple channels at the same time, without any need of additional radiation power and

additional spectrum bandwidth utilization [6]. For antenna designers, it is very challenging to design a MIMO antenna with multiple patch elements on the same substrate [7], i.e. to achieve proper spacing between the individual elements in order to avoid mutual coupling [8], and to ensure good isolation within the wide frequency band.

In this paper, a compact fractal MIMO antenna for UWB applications is designed with a metamaterial structure serving as the ground plane. Fractal geometry is used for miniaturization purposes and to increase the perimeter of the patch. The metamaterial structure helps in enhancing the bandwidth of the antenna and in boosting performance in terms of impedance matching, bandwidth, and peak gain. Because of the maximum miniaturization potential, the spaces between radiating components are very small, and coupling between them is calculated to achieve a good diversity performance attained due to the use of the metamaterial structure. The design is worked out and optimized with the use of HFSS software.

2. Antenna Geometry

The design is based on a typical FR4 substrate material. The design process incorporates four iterations to obtain the correct fractal structure (Fig. 1). The process begins with a monopole antenna using a hexagonal slot. In the course of the iterations, another hexagonal slot is loaded into the previous fractal iteration and so on. The design process is limited to four iterations due to fabrication-related issues.

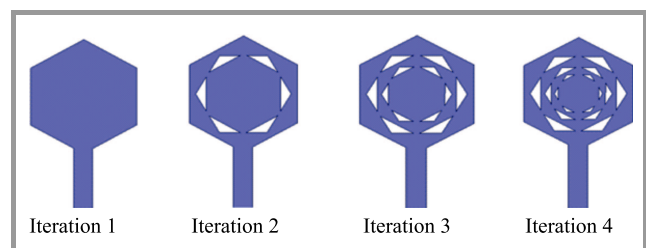


Fig. 1. Fractal iterations of the antenna design.

By using a circular monopole, the radius of the hexagonal patch is derived from:

$$R = \frac{F}{\left\{1 + \frac{2h}{\pi\epsilon_r F} \left[\ln \frac{\pi F}{2h} + 1.7726\right]\right\}^{\frac{1}{2}}}, \quad (1)$$

where:

$$F = \frac{8.791 \cdot 10^9}{f_r \sqrt{\epsilon_r}}. \quad (2)$$

The final MIMO antenna is shown in Fig. 2. The bottom layer of the antenna contains the metamaterial split ring resonator structure (Fig. 3), comprising two rectangular rings with a split on either side of the rings. This structure is developed by removing the inner middle part of the ground and inserting another rectangular ring inside the outer ring. Together, the inner ring and the outer ring form the split ring structure.

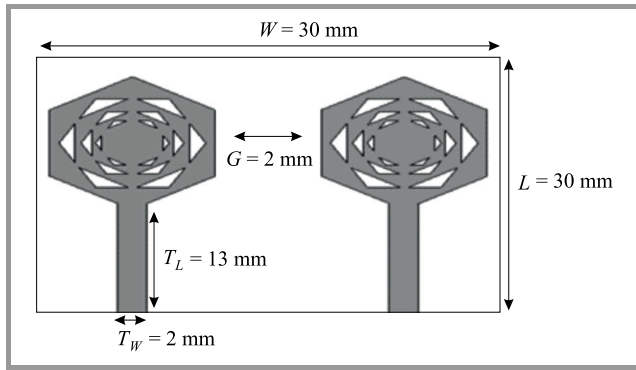


Fig. 2. Final MIMO antenna design.

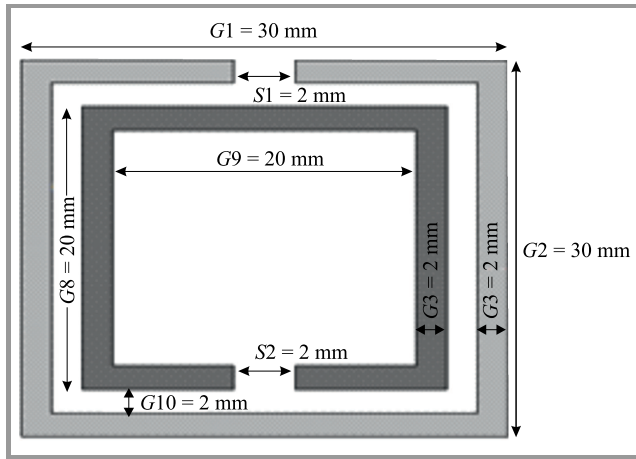


Fig. 3. Split ring resonator on the lower side of the substrate.

3. Results and Discussions

The comparison of the return loss parameter for all four fractal iterations is shown in Fig. 4. The minimum return loss value is achieved in the last iteration. Upon comparing the outcomes, it becomes obvious that as the iteration number increases, a good impedance matching is observed

between the feed line and the patch. Simulation results show that the operating band of 5.8–15 GHz is obtained.

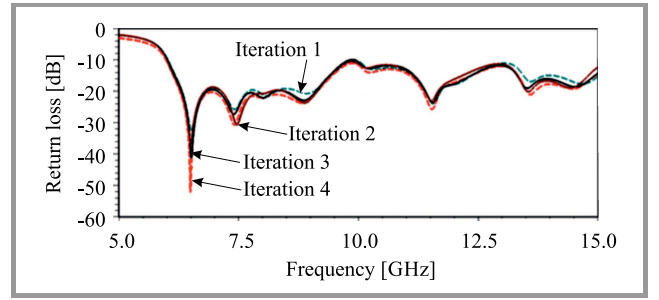


Fig. 4. Return loss comparison of each iteration. (For color pictures see the digital version of the paper).

Figure 5 shows the simulated gain of the proposed antenna for all four iterations. Positive gain is observed throughout the entire operating band. The highest gain of 6.2 dB is observed at 14 GHz. It is also noticed that there is not much difference in the gain as the iteration number increases.

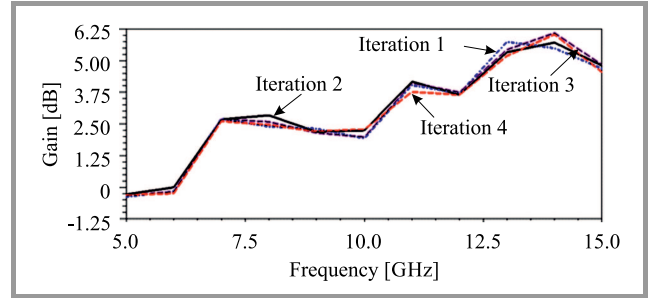


Fig. 5. Frequency vs. gain of the proposed antenna.

Figure 6 presents radiation patterns at E plane and H plane at 5, 8 and 12 GHz, respectively. It shows that the H plane pattern is omnidirectional and that the E plane pattern is bidirectional, therefore forming a stable pattern.

The fabricated prototype of the proposed antenna is shown in Fig. 7. It has the overall dimensions of 30×30×1.6 mm. The gap between two radiators is just 2 mm. A bigger gap between the radiators might lead to an increase in overall antenna dimensions. The antenna is designed in such a way that good isolation is ensured between the radiating elements, despite the fact that they are placed in close proximity to each other.

Figure 8 illustrates the measured performance of the MIMO antenna that is compared with the simulation data. Both the simulated and measured parameters are nearly the same and prove that the solution is suitable for UWB applications. Figure 9 shows the measured and simulated mutual coupling characteristics of the antenna. The measurement shows that the coupling between the two elements is below −25 dB. In addition, it was observed that the use of SRR structure reduces the coupling from −10 dB to −25 dB. In Fig. 10, the measured and the simulated peak gain is illustrated for different frequencies.

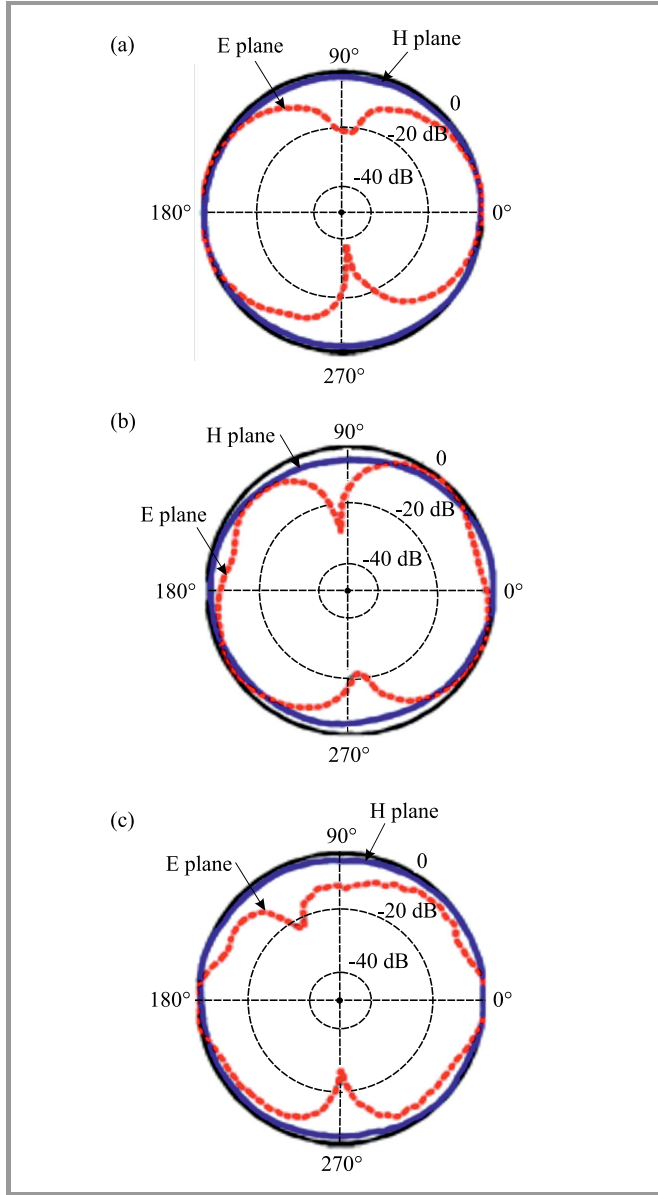


Fig. 6. Measured radiation patterns at: (a) 5 GHz, (b) 8 GHz, and (c) 12 GHz of (blue line: E plane and red (dashed) line: E plane).

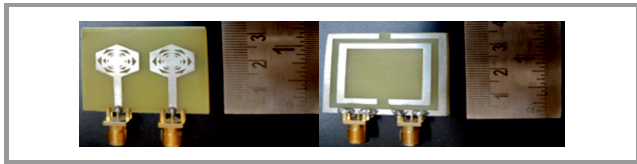


Fig. 7. Fabricated prototype of the antenna.

Figure 11 presents the distribution of surface current on the radiating elements and on the metamaterial structure at 3 frequencies: 5, 8 and 12 GHz. This pattern is obtained using the HFSS simulator. The magnitude of the antenna's surface current is displayed in different colors, and the magnitude levels with their corresponding colors are shown in the legend, next to each diagram. The blue color at the bot-

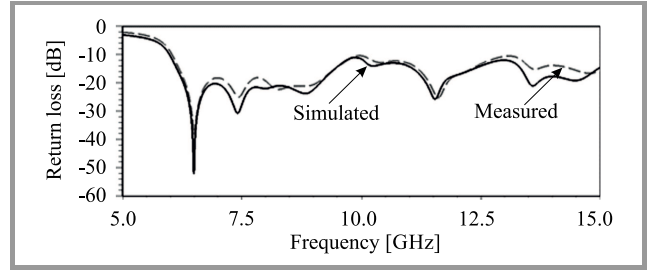


Fig. 8. Simulated vs. measured return loss of the proposed antenna.

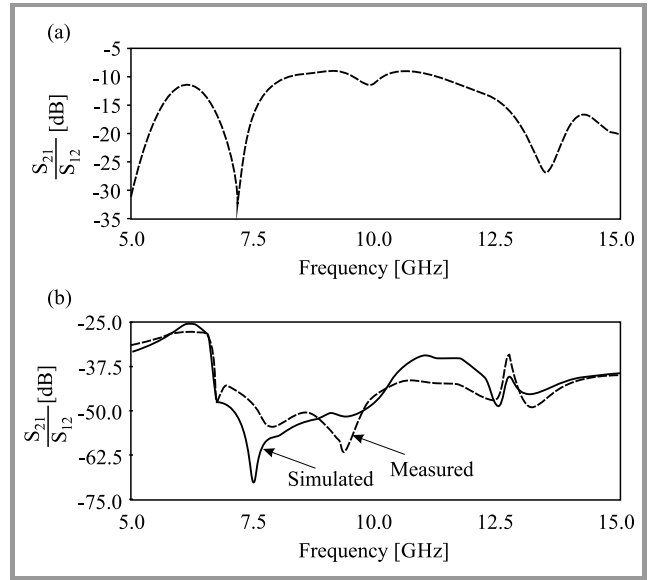


Fig. 9. Isolation characteristics between the radiating elements: (a) without SRR and (b) with SRR.

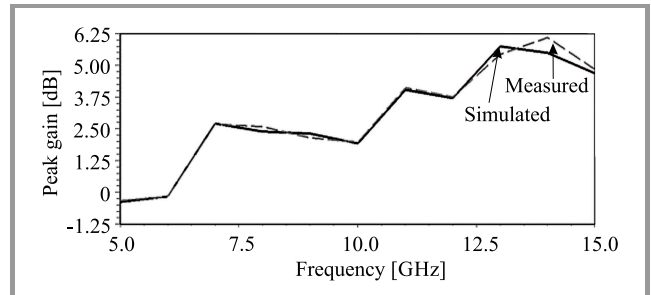


Fig. 10. Peak gain vs. frequency of antenna.

tom of the legend box correspond to the minimum surface current and the red color shows the maximum current. In Fig. 10a, the fractal portion is yellow in color at 5 GHz indicating high current distribution, whereas in Fig. 10c, the fractal portion is blue in color at 12 GHz, indicating relatively low surface current distribution. Therefore, current intensity is high at low frequencies, in comparison to high frequencies.

The capability (diversity performance) of a MIMO antenna may be validated using the ECC parameter. Based on S parameters, ECC may be calculated using [15]:

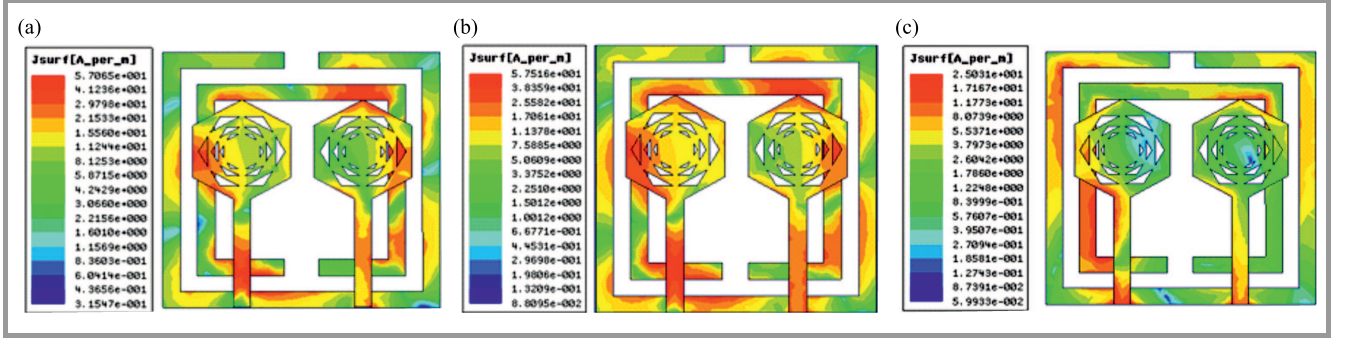


Fig. 11. Current distribution at: (a) 5 GHz, (b) 8 GHz, and (c) 12 GHz. (For color pictures see the digital version of the paper).

$$ECC = \frac{|S_{11}^* S_{12} + S_{21}^* S_{22}|^2}{(1 - |S_{11}|^2 - |S_{21}|^2)(1 - |S_{22}|^2 - |S_{12}|^2)}, \quad (3)$$

Ideally, the ECC value should be zero, but the typical limit is <0.5 . For the desired antenna, the ECC is below 0.025 over the frequency range of 6 to 15 GHz, as shown in Fig. 12 and remains within the aforementioned limit.

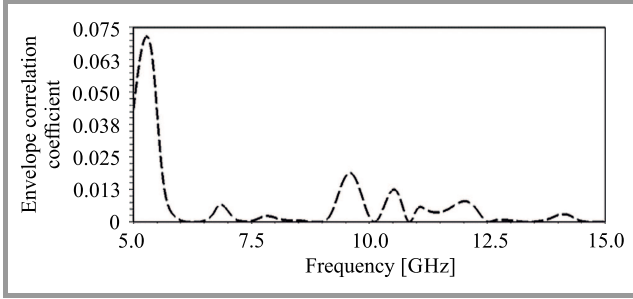


Fig. 12. ECC plot.

The effectiveness of diversity can be assured based on another important parameter, known as diversity gain (DG). DG is a function of ECC and may be calculated using [15]:

$$DG = 10 \cdot \sqrt{1 - |0.99 \cdot ECC|^2}. \quad (4)$$

DG of the desired antenna is illustrated in Fig. 13 and equals almost 10 dB.

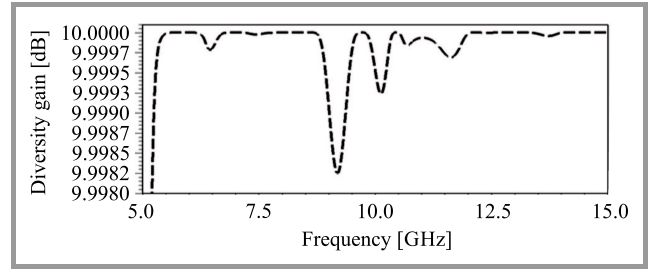


Fig. 13. Diversity gain.

The comparison of performance the proposed antenna with other designs known from literature is shown in Table 1. The proposed solution is compact compared to other designs and supports the maximum frequency range of 15 GHz. While the percentage bandwidth may be higher in a few other papers, the proposed design offers the best isolation of -25 dB with good ECC and diversity gain. The spacing between two radiating elements is just 3 mm. When compared with other available designs the coupling between the elements is reduced significantly, despite the fact that the elements are placed very close to each other.

Table 1
Comparison of performance of the designed antenna
and of other solutions proposed in literature

Reference	Dimensions [mm]	Frequency band [GHz]	Bandwidth percentage	Isolation [dB]	ECC	DG [dB]
[9]	45×37	3.1–5	47	< -12	<0.5	NA
[10]	50×30	2.5–14.5	141.17	< -17	<0.04	> 7.4
[11]	40×80	4.5–8	56	< -20	<0.02	> 9.94
[12]	35×68	3.1–10.65	109.89	< -16	<0.04	NA
[13]	40×40	3.1–11	112	< -22	<0.02	> 9.98
[14]	48×48	3–11	114	< -17	NA	NA
Proposed solution	30×30	5.8–15	90	< -25	< 0.025	> 9.9

4. Conclusion

A compact MIMO antenna is proposed and analyzed in this paper for four fractal design iterations. The obtained bandwidth is 9.2 GHz with a maximum gain of 6.2 dB at 14 GHz and with a reduced mutual coupling of -25 dB throughout the operational frequency. The antenna has a bidirectional pattern of radiation in the H plane and an omnidirectional pattern along the E plane. The stable radiation patterns, coupling characteristics and reflection coefficients show that the antenna may be used for ultra-wide band applications. Also, the obtained ECC and DG values prove good diversity performance. The impedance bandwidth of this MIMO antenna is suitable for WiMAX and WLAN applications as well.

References

- [1] A. Mehdipour, A. R. Sebak, and C. W. Trueman, "Novel ultrawideband slot-antenna with dual band-notched characteristics", in *IEEE Antenn. and Propag. Society Int. Symp.*, San Diego, CA, USA, 2008 (DOI: 10.1109/APS.2008.4620056).
- [2] M. A. Matin, "Stacked E-Shaped patch antenna for lower-band ultrawideband (UWB) applications", in *IET Int. Conf. on Wireless, Mobile and Multimedia Networks*, Beijing, China, 2008 (DOI: 10.1049/cp:20080132).
- [3] M. H. Reddy, R. Sheeba, and S. T. Niba, "Design of microstrip patch antenna for space environment with minimum return loss", *Res. J. of Pharmaceutical, Biological and Chemical Sci.*, vol. 8, no. 2, pp. 2678, 2017 [Online]. Available: [https://www.rjpbcs.com/pdf/2017_8\(2\)/\[300\].pdf](https://www.rjpbcs.com/pdf/2017_8(2)/[300].pdf)
- [4] M. H. Reddy, R. M. Joany, G. Manikandan, and A. S. A. Nisha, "Design of microstrip patch antenna with multiple slots for satellite communication", in *Int. Conf. on Commun. and Signal Proces.*, April 2017, Chennai, India, 2018 (DOI: 10.1109/ICCSP.2017.8286482).
- [5] M. Sugadev and E. Logashanmugam, "Quad band U-slot microstrip antenna for C, X and Ku band wireless applications", *Contemporary Engin. Sci.*, vol. 8, no. 13, pp. 737–745, 2015 (DOI: 10.12988/ces.2015.53115).
- [6] W. P. Siri Wongpairat and K. J. R. Liu, *Ultra-Wideband Communications Systems: Multiband OFDM Approach*. Wiley-IEEE Press, 2007 (ISBN: 9780470074695).
- [7] Y. Rahayu, "Reconfigurable ultra wideband antenna design and development for wireless communication", Ph.D. Thesis (unpublished), Universiti Teknologi Malaysia, Malaysia, 2008.
- [8] Y. Zehforoosh, C. Ghobadi, and J. Nourinia, "Antenna design for ultra wideband application using a new multilayer structure", *PIERS Online*, vol. 2, no. 6, 2006 (DOI: 10.2529/PIERS060531145356).
- [9] T. S. P. See and Z. N. Chen, "An ultrawideband diversity antenna", *IEEE Trans. on Antenn. and Propag.*, vol. 57, no. 6, pp. 1597–1605, 2009 (DOI: 10.1109/TAP.2009.2019908).
- [10] A. Iqbal, O. A. Saraereh, A. W. Ahmad, and S. Bashir, "Mutual coupling reduction using F-shaped stubs in UWB-MIMO antenna", *IEEE Access*, vol. 6, pp. 2755–2759, 2017 (DOI: 10.1109/ACCESS.2017.2785232).
- [11] A. H. Jabire, H.-X. Zheng, A. Abdu, and Z. Song, "Characteristic mode analysis and design of wide band MIMO antenna consisting of metamaterial unit cell", *Electronics*, vol. 8, no. 1, 2019 (DOI: 10.3390/electronics8010068).
- [12] W. T. Li, Y. Q. Hei, H. Subbaraman, X. W. Shi, and R. T. Chen, "Novel printed filtenna with dual notches and good out-of-band characteristics for UWB-MIMO applications", *IEEE Microw. and Wirel. Compon. Lett.*, vol. 26, no. 10, pp. 765–767, 2016 (DOI: 10.1109/LMWC.2016.2601298).
- [13] W. A. E. Ali, A. A. Ibrahim, "A compact double-sided MIMO antenna with an improved isolation for UWB applications", *AEU – Int. J. of Electronics and Commun.*, vol. 82, pp. 7–13, 2017 (DOI: 10.1016/j.aeue.2017.07.031).
- [14] C.-X. Mao and Q.-X. Chu, "Compact coradiator UWB-MIMO antenna with dual polarization", *IEEE Trans. on Antenn. and Propag.*, vol. 62, no. 9, pp. 4474–4480, 2014 (DOI: 10.1109/TAP.2014.2333066).
- [15] M. H. Reddy, D. Sheela, V. K. Parbot, and A. Sharma, "A compact metamaterial inspired UWB-MIMO fractal antenna with reduced mutual coupling", *Microsystem Technologies*, pp. 1–13, 2020 (DOI: 10.1007/s00542-020-05024-z).



Mekala Harinath Reddy received his undergraduate degree in Electronics and Communication Engineering from Sathyabama University, Chennai, India in 2017. He received the Star of Excellence Award for the best graduate if Sathyabama University. His interests are in microwaves and antennas. He has published more than 8 papers in indexed

international journals and conferences. Currently, he is pursuing an M.Sc. in Information and Communications Engineering at the University of Leicester, UK.

 <https://orcid.org/0000-0003-0321-5421>

E-mail: harinath.m1995@gmail.com

School of Engineering
University of Leicester
Leicester
United Kingdom



D. Sheela received her Ph.D. degree in Information and Communication Engineering from Anna University in Chennai and is working as a Professor in the Department of Electronics and Communication Engineering, Tagore Engineering College, Chennai. She has 28 years of teaching, research and industry experience. She published several papers in reputed inter-

national journals and conferences. Her areas of interests are optical communication, networking, wireless communication and antennas.

 <https://orcid.org/0000-0002-0974-3922>

E-mail: dsheelatagorerandd@gmail.com

Department of ECE
Tagore Engineering College
Chennai
India

A Comprehensive Survey on Resource Management in Internet of Things

Lokesh B. Bhajantri¹ and Gangadharaiah S.²

¹ Department of Information Science and Engineering, Basaveshwar Engineering College, Bagalkot, India

² Department of Computer Science, S.R.S.M.N. Government First Grade College, Barkur, India

<https://doi.org/10.26636/jtit.2020.145220>

Abstract—Efficient resource management is a challenging task in distributed systems, such as the Internet of Things, fog, edge, and cloud computing. In this work, we present a broad overview of the Internet of Things ecosystem and of the challenges related to managing its resources. We also investigate the need for efficient resource management and the guidelines given/suggested by Standard Development Organizations. Additionally, this paper contains a comprehensive survey of the individual phases of resource management processes, focusing on resource modeling, resource discovery, resource estimation, and resource allocation approaches based on performance parameters or metrics, as well as on architecture types. This paper presents also the architecture of a generic resource management enabler. Furthermore, we present open issues concerning resource management, pointing out the directions of future research related to the Internet of Things.

Keywords—Internet of Things, resource allocation, resource discovery, resource management.

1. Introduction

The Internet of Things (IoT) connects physical world objects around us to the Internet. The heart of the IoT ecosystem comprises a huge set of smart devices with sensors and actuators, offering also a certain amount of computational and communication capabilities [1]. The operating principle of IoT devices or a system of smart objects is shown in Fig. 1. The system performs a number of tasks, such as sensing, communication, computation and actuation, which are required for collecting and processing data within the IoT environment. The sensors used may be of different types and may measure, for instance, temperature, humidity, pressure, etc. All such devices are used to gather data from their surrounding environments. In the next step, this information is uploaded to the server, either directly or via a gateway. Based on server feedback, IoT devices may send commands to actuators in order to properly adjust the specific parameters.

IoT finds numerous applications in the real-world environment, e.g. in smart homes, intelligent transportation, health

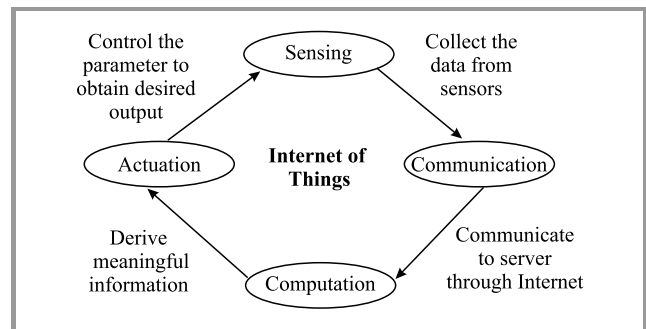


Fig. 1. Operating principle of IoT devices or smart objects.

monitoring, retail activities, smart cities, environment monitoring, power management, etc. [2], as shown in Fig. 2.

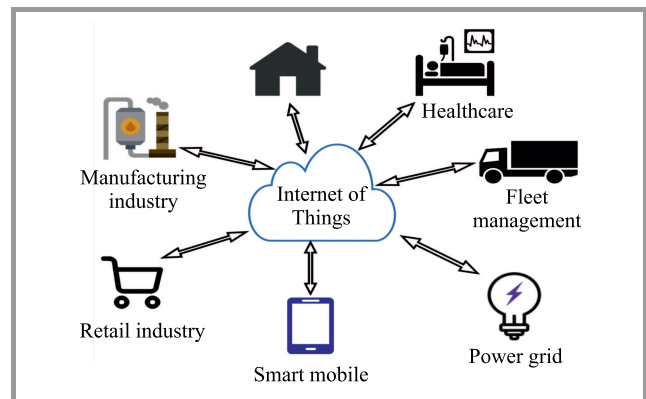


Fig. 2. Applications of IoT.

Due to the considerable benefits offered by IoT applications, industries, R&D labs and governments from across the world are investing huge amounts in IoT technologies. This has led to fragmentation and strong competition on the IoT market that suffers from incompatibility due to the existence of multiple standards. The International Telecommunication Union for Telecommunication (ITU-T) has recognized the different challenges faced by IoT networks in comparison to traditional networks. ITU-T has proposed

a new reference model for the IoT environment [3]. It consists of four layers: device layer, network layer, service support layer and application support layer. Management and security capabilities are taken into consideration as well. Since management and security capabilities apply to all four layers, they are represented vertically in the ITU's reference model.

The new paradigm of connected smart devices creates a number of research challenges, related mainly to interoperability across heterogeneous networks, machine-to-machine communication, self-aware and self-organizing networks, open framework for the IoT, large-scale deployment of infrastructure, dynamic, autonomous and adaptive resource management, ad-hoc deployable and configurable networks, distributed energy efficient data processing, mechanisms protecting against attacks, access control and accounting schemes for IoT, lightweight secure and high performance authentication, standardization of APIs, and context-aware adaptation of operation [4].

Traditional computer network architectures and reference models may not be suitable to tackle these challenges and issues. Many researchers have suggested an increased number of layers compared to TCP/IP, but none is standardized in the IoT environment (Fig. 3). The conventional three-layer architecture comprises application, network, and perception layers, whereas in the five-layer architecture, business, application, processing, transport and perception layers are distinguished.

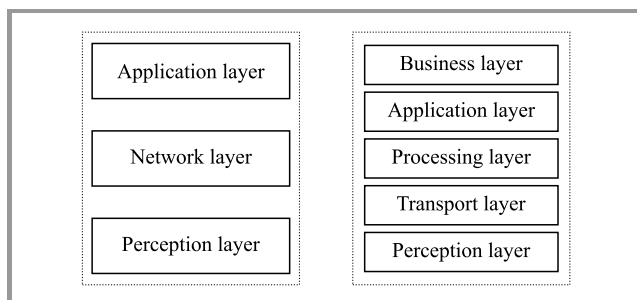


Fig. 3. Three- and five-layer architectures.

The application layer used in both architectures provides user specific services, such as health monitoring systems, fleet management, environment monitoring, and so on. The perception layer is responsible for managing sensors and actuators. It gathers such information as temperature, humidity, motion, etc., and forwards it to the network layer (in the 3-layer architecture) or to the transport layer (in the 5-layer architecture). It is responsible for networking the functionalities of sensing devices, gateways and the server. In the five-layer architecture, transport layer is responsible for transferring data using wireless technologies, e.g. cellular, Wi-Fi or Bluetooth, from the perception to the processing layer and vice versa. The processing layer is responsible for storing and analyzing data collected from the transport layer. It uses other technologies, such as cloud, fog or edge computing for processing such data. The busi-

ness layer manages the entire IoT ecosystem in order to provide services to the user.

In this paper, we focus on existing solutions and open issues related to managing resources in IoT ecosystems, i.e. on the basic functionality of IoT that needs to be supported in accordance with the ITU reference model. Research on IoT resource management is scarce, but the works available tend to focus on the stages of resource management processes, as well as on the classification of resource allocation methods based on energy, context, quality of service (QoS), service level agreement (SLA), load balancing, cost, efficiency, etc. [5]. WSN-based IoT networks, resource types and resource scheduler methods are taken into consideration as well, just as are resource allocation algorithms, integrated approaches to IoT and fog networks, as well as IoT, fog and edge devices.

But these works fail to discuss the overall nature of the process, e.g. what resources are required to perform the allocation. How it is modeled? Which types of infrastructures are used? How to take advantages of edge or fog computing in delay-sensitive applications? IoT networks integrate multiple heterogeneous networks, such as personal area network (PAN), mobile edge computing, and industrial IoT, thus creating new challenges. Information is lacking on how resources are identified, modeled and selected before the scheduling process. Taking into consideration the shortcomings of work that has already been performed, we have presented, in a systematic manner, the process of managing resources in IoT, the potential options available for modeling, and the existing algorithmic solutions. Finally, we also outline research-related challenges and open issues concerning IoT resource management.

In this work, we outline the importance of each phase of the resource management process and describe their interconnection with the reference model and SDOs. The main focus is placed on resource allocation algorithms and the link between this stage and the remaining phases of the process. The following are the most significant contributions of our work:

- systematic study of multi-tier architectures (and their stages) used for resource management,
- understandings different SDO and industry guidelines suggested for managing IoT resources,
- outline of existing resource modeling, resource discovery, resource estimation and monitoring techniques relied upon in IoT resource management,
- analysis of existing resource allocation approaches based on performance metrics and architectures,
- existing issues and future trends in IoT resource management.

The remaining part of this paper is organized as follows: Section 2 presents an overview of resource management

in IoT. The IoT resource management module used in Fiware is described in Section 3. Challenges and research-related issues concerning IoT resource management are outlined in Section 4. Finally, the discussion is concluded in Section 5.

2. Resource Management in IoT

This section describes the ITU-T reference model for IoT, ecosystems and the individual resource management stages. In an IoT ecosystem, the participating entities are heterogeneous in nature and rely on different processing capacities (8/16/32/64-bit), communication technologies (cellular/Bluetooth/Wi-Fi/ZigBee/Rfid/LoRa) and storage capacities. The sensing devices may be deployed in a deep forest, may be a part of energy distribution systems used in cities, or may be installed in moving vehicles.

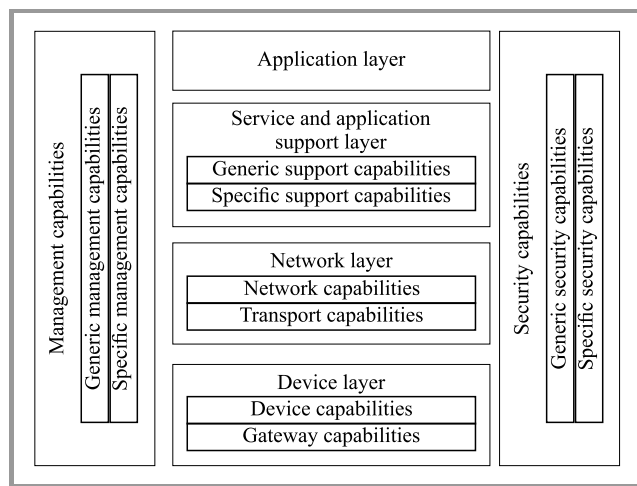


Fig. 4. ITU-T reference model for IoT.

Each IoT application or service – e.g. real time video streaming or simple event notifications – may have different requirements. Deploying, maintaining and monitoring a wide range of devices is a tedious task in a real world environment. In order to address heterogeneity and complexities of the network, ITU-T has proposed an IoT reference model shown in Fig. 4 [3]. The description of each layer is presented below:

- **Application layer** – comprises with end user IoT applications.
- **Service support and application support layer** – offers generic and specific support capabilities. Generic capabilities are used to offer common functionalities that may be used by any specific applications. Specific support capabilities provide services that are accessed by specific applications.
- **Network layer** – supports network transport capabilities. As far as network capabilities are concerned,

it supports device connectivity, mobility management, routing, authorization and authentication for diverse protocols. In terms of transport capabilities, it supports the transportation of application specific data, IoT services, control and management information.

- **Device layer** – this layer supports device and gateway capabilities. Device capabilities include direct interaction with the communication network, indirect interaction with the communication network, ad-hoc networking between devices, as well as sleep and wake-up cycle for energy conservation. Gateway capabilities include protocol conversion and multi interface support.
- **Security capabilities** – support for application specific and generic capabilities. General capabilities include authorization, authentication, user privacy protection, security audits in the application layer, secure communication, data and signaling at network layer, as well as authentication, authorization and access control over the device layer.
- **Management capabilities** – generic and specific capabilities are distinguished here. Specific capabilities deal with application specific requirements. Generic capabilities, in turn, perform functions for specific applications, including device activation, deactivation, remote access, device status, software updates, topology and traffic management. The management layer is responsible for fault detection, accounting, configuration, security and for performance of device and software components.

2.1. Resources in IoT Ecosystems

The various types of resources used in IoT environments are shown in Fig. 5. The resources may perform any tasks that are useful for the system, network or end user applications. In general, there are two main types of resource management approaches that are concerned with infrastructure and applications. Infrastructure-based resource management relates to computational, networking, storage and energy resource managements. From the system perspective, hardware (sensor, CPU, memory), firmware and software resources may be distinguished. From the network prospective, the resources may include the radio antenna, channel, bandwidth, routing path or nodes. Storage resource management is associated with various types of different components, such as memory, file system and so on. Similarly, energy resource management deals with energy, battery and so on. From the application or user perspective, resource management may affect any software components, such as application management module, resource information base, customer information base, resource identification or resource modeling.

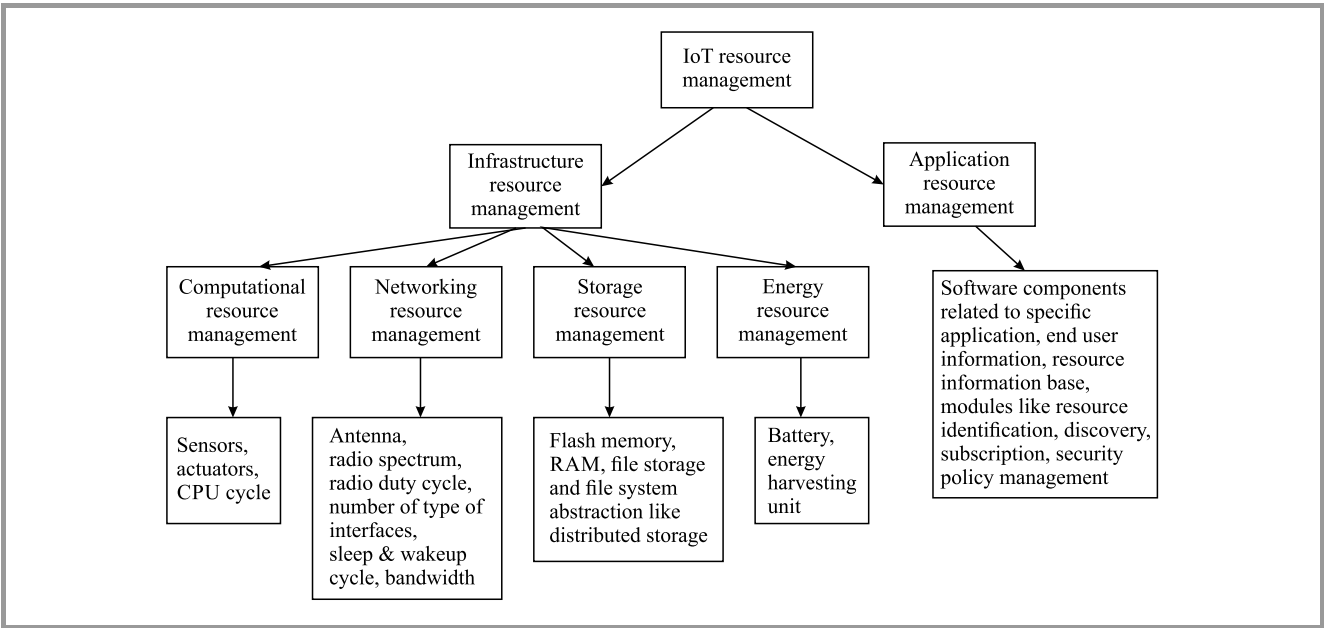


Fig. 5. Types of IoT resources.

2.2. Distributed Architecture of IoT Ecosystems

Data collected by sensor nodes has to be processed before being used for the provision of meaningful services to users or applications. Processing consumes multiple re-

sources that are geographically distributed. It involves heterogeneous hardware, software and communication technologies. Figure 6 shows the typical operating principle of a distributed IoT ecosystem.

Sensing devices may be IP or non-IP based. Non-IP based IoT devices are generally low powered devices using such protocols as Bluetooth, ZigBee and Wi-Fi. These devices forward the collected information to the nearest configured gateway. Data collected at each gateway is filtered and processed before being forwarded to the central server or cloud storage. IP-based IoT devices will forward the data either through the gateway, or they may upload it directly to the central server. Distributed architecture is associated with the following computing mechanisms that are affect resources in the IoT environment:

- **Cloud computing resources in IoT ecosystems.** Clouds rely on vast computational and storage resources. IoT devices may be used in cloud computing infrastructures for storing and performing a complex analysis of IoT data. Cloud computing offers reduced storage, computational and operational cost. But the integration of IoT and cloud gives rise to many challenges [6]. Seamless Internet connectivity is required, mainly, and latency between cloud and IoT devices is also very high, which is unacceptable for real-time applications, such as health monitoring systems. Additionally, security and privacy will also pose major challenges, since data is stored remotely will also be a major challenge since data is stored remotely.
- **Edge computing resources IoT ecosystems.** Edge computing offloads the resource demand from end devices and cloud. Edge devices may include the

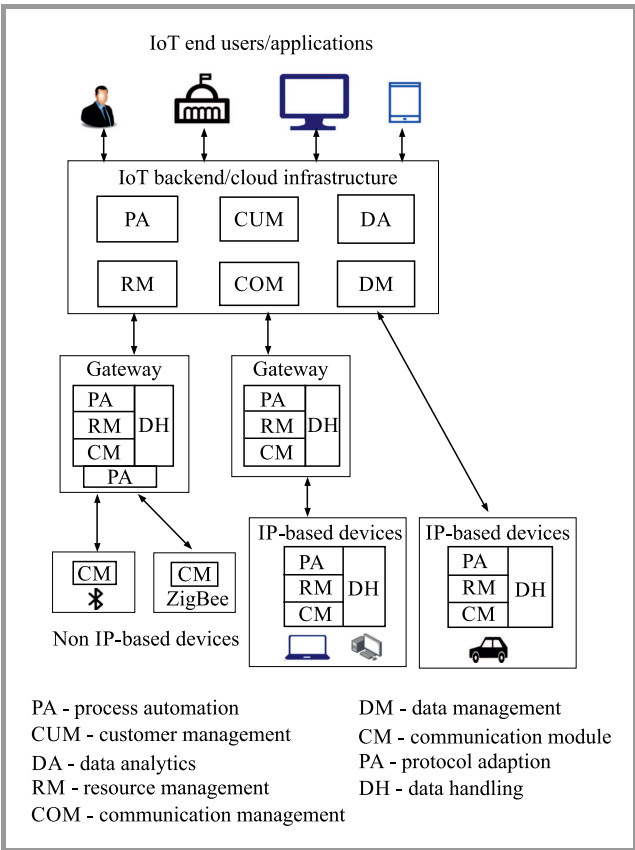


Fig. 6. Distributed architecture of IoT system.

gateway, device controller or smartphone. Real-time applications that demand low latency and high response times benefit from edge computing. Edge computing contributes also to efficient context-aware computing and to mobility management.

- **Fog computing resources in IoT ecosystems.** The fog computing paradigm is similar to that of edge computing, bringing computation, storage and network services near the data source, instead of transferring them to cloud computing. But in fog computing, one or more devices in LAN help in offloading the workload of edge devices. This approach offers advantages that are similar to those of edge computing, but with a much higher processing and storage capacity.

Resource requirements of distributed IoT systems depend solely on the system's functional and non-functional requirements. Detailed specifications are provided by ITU-T [7]. The different requirements and their classification are presented below:

- business requirement specification – a high-level document that specifies the organization's motivation to development of a new system,
- stakeholder requirement specification – includes the perspective and requirements of different stakeholders, such as users/operators/maintenance personnel, related to using the system to generate business contributions,
- system requirements specification – specifies the technical requirements of the selected system, including those of functional and non-functional character:
 - **functional requirement specification** – specifies the behavior of the system in terms of service/function/operation. The important functional requirements specified by ITU for IoT networks include application support requirements, service requirements, communication requirements, device requirements, data management requirements, security and privacy protection requirements,
 - **non-functional requirement specification** – specifies supplementary/quality-related characteristics of the system. The important non-functional requirements specified by ITU-T for IoT networks include interoperability, scalability, reliability, high availability, adaptability and manageability. Non-functional resource requirements are difficult to calculate, as they are of the subjective and relative nature [8]. An attempt to improve one parameter may exert a direct impact on other parameters of the same system.

In the IoT environment, management of resources is a crucial task in terms of fault tolerance and security. The resource management functionality is distributed across the different layers of the reference model. Any hardware, software and communication modules can be modeled as resources.

2.3. Resource Management in IoT

The individual stages of the resource management process are shown in Fig. 7. Resource modeling is an important stage in resource management. It is the process of abstracting and building metadata for IoT resources. Effective resource modeling is the key to efficient resource discovery and allocation mechanisms in IoT networks. Resources are modeled based on semantic, virtualization and attribute modeling methods. Each of these modeling methods is described below:

- **Semantic modeling.** This model helps build loosely coupled, interoperable and service-oriented IoT architectures. Device capabilities or services rely on standard or machine-readable formats, which enhances machine-to-machine (M2M) communication, derivation of new knowledge and interoperability [9]. Different SDOs are working to address issues related to standardization. Sensor web enablement by Open Geospatial Consortium (OGC) addresses mainly sensors, sensor data models and sensor web services [10], [11]. It also includes a specification related to observation and measurement (O&M), sensor modeling language (SML), transducer model language (TML), sensor observation services (SOS), etc. W3C defines also the formal specification for data representation by means of a resource description framework (RDF) and ontology representation language, such as ontology web language (OWL).
- **Virtualization based modeling.** In this model, sensor hardware, software or network resources are represented virtually in a cloud or a gateway, for ease of management and for optimized resource utilization [12]. Different types of virtualizations relied upon in IoT are listed below:
 - **Software defined network (SDN).** SDN abstracts the physical network, which helps in controlling and configuring the network behavior centrally. SDN has two planes – one is the forward plane for packet transmission, and the other is the control plane for network configuration and management. Decoupling these planes aids in element of heterogeneous devices;
 - **Network function virtualization (NFV).** NFV eliminates the need for dedicated hardware and uses a commodity server for network functions. Different network functions, such as routing,

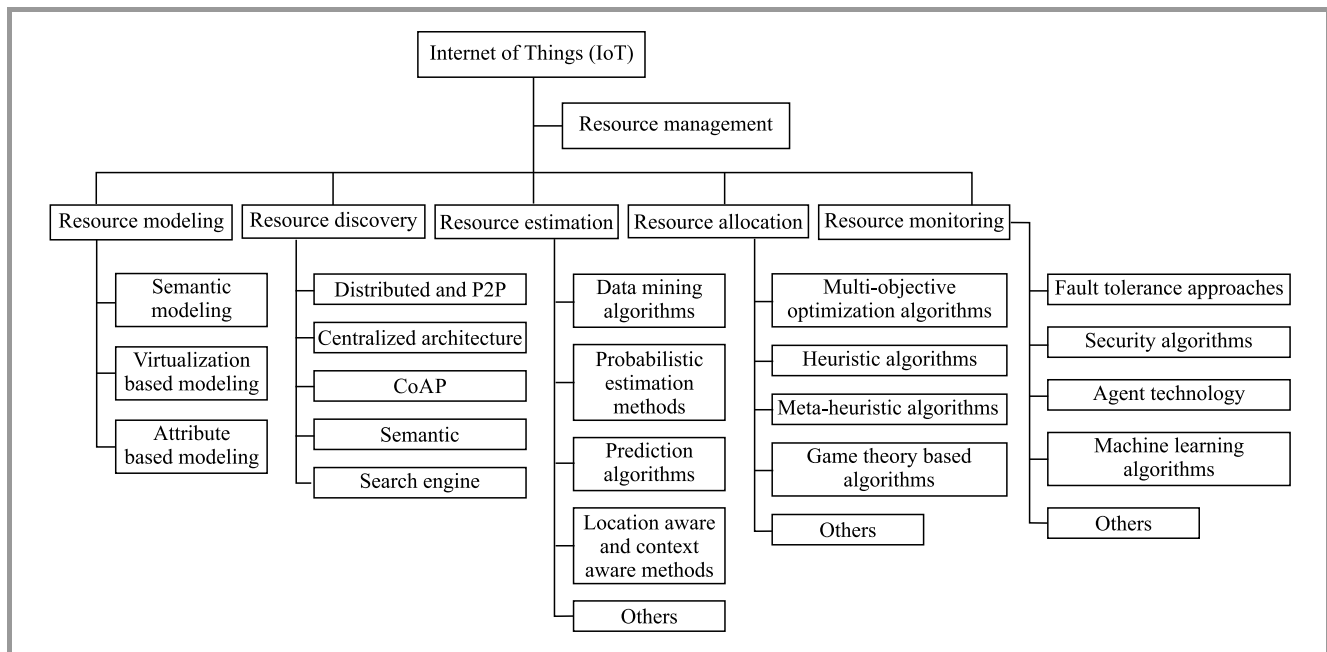


Fig. 7. Stages of the resource management process.

firewall, load balancing and intrusion prevention may be performed by the commodity server due to NFV implementation;

- **Containers/dockers.** This approach is an OS level virtualization, where applications are packed with their own dependencies, libraries and configuration files. Since multiple containers are executed directly on the host OS, the solution is more lightweight than virtual machines. In IoT ecosystems, many cloud service providers support development and management of containers [13]. Containers are generally deployed in edge devices, helping in real time data analysis (e.g. packet inspection and intelligent alert systems). They also act as load balancers by reducing the load of cloud devices;
- **Sensor virtualization.** In this approach, an abstract model of IoT devices is hosted on the central server. End user applications submit requests to the central server, with an abstract IoT device representation, rather than to physical IoT devices. Sensor virtualization enhances energy efficiency and lowers maintenance cost [14].

- **Attribute based modeling.** In attribute based resource modeling, resources are designed based on functionality and attributes of devices, such as their location, energy, sensing parameters, etc.

Results of some research projects related to resource modeling have been published recently. In [15], the authors suggest a set of tasks to be performed by a set nodes in polynomial time. While modeling the metadata, the author

has considered various properties and constraints of nodes. Such constraints as location, sensing capacity, power, software and hardware are modeled in the form of a binary programming model and are then applied using the heuristic method. In paper [16], the author modeled the problem of assigning different IoT services to different physical interfaces. In this approach, interfaces are modeled as resources. Since the interfaces are limited in any system, multiple services compete for those interfaces that are available. The author proposes two algorithms: single round assignment – when sufficient resources are available, and multiple rounds – when resources cannot be shared in the single round. Service splitting is performed to handle the workload. Splitting and distribution cost is also considered. Another approach is based on application demand and resources are identified based on the deployed WSNs [17]. Then, the linear programming model is applied to map the resources and application requirements. It considers different attributes, such as functional capability of the node, granularity of operation, power consumption, location, duty cycle and fidelity region. In [18], the authors discuss hardware, software and communication resources that are integrated to provide services (i.e. everything as a service) to IoT user. They also place a greater emphasis on information rather than on infrastructure.

The work presented in [19] has led to designing an IoT service discovery and ranking method that uses semantic modeling. Ontology linking is used to create semantic reasoning for better IoT services. Semantic based resource modeling approaches are effective in handling heterogeneous, distributed, and ultra large-scale resources in IoT. Paper [20] presents the OpenIoT platform that uses W3C SSN ontology. It also explains the process of implementing IoT applications with cloud integration. The W3C SSN on-

Table 1
Comparison of different resource modeling techniques

Refer-ences	Attribute-based approach	Semantic-based approach	Virtualiza-tion-based approach	Method used	Proposed approach
[15]	●	○	○	Binary programming model is developed to represent attributes of IoT device and its constraints.	Heuristic method to improvise execution time and load balance.
[16]	●	○	○	Mathematical model considers services and interfaces as resources.	Mapping of interfaces to services improves the utilization, service, and activation cost.
[17]	●	○	○	Deployed WSNs resources are modeled as a resource map for satisfaction of IoT requirements.	The framework optimizes the resource mapping and allocation time.
[18]	○	●	○	Hardware, software and network are modeled as a service.	A new service model where hardware, software and communication modules are integrated as a single service for the user.
[19]	○	●	○	Ontology linking is performed for service discovery and ranking.	Scalable solution based on semantic reasoning.
[20]	○	●	●	SSN ontology is used for middleware, supporting virtual sensors in cloud.	The model is based on the W3C SSN ontology which provides better interoperability.
[21]	○	●	○	A new, SSN ontology-based design pattern is proposed.	Lightweight semantic modeling method, extendable due to linked data.
[22]	●	●	○	NFV is applied at intermediate nodes.	Two intermediate processing nodes for better execution time and response time.

tology is described from the perspective of sensors, sensing capabilities, stimuli and observations [21]. The work proposes also the stimulus-sensor-observation ontology design pattern, which relies on light-weight semantics preferred by several linked data applications.

The importance of semantic interoperability and its challenges are presented in [22]. The author discusses about challenges related to data modeling, information exchange, semantic annotation and semantic discovery of data sources, semantic reasoning and interpretation, as well as knowledge representations. The work outlines also a few important requirements for IoT semantic platforms. In paper [23], the author quantitatively evaluated four different IoT middleware platforms using a small-scale IoT scenario. In this work, a comparison of functional components, such as service registration, discovery and composition, was performed. The work highlights the need for an autonomous system with automatic service registration, discovery and composition to handle heterogeneous devices in IoT environments.

Virtualization of IoT devices for such applications as smart cities is discussed in [24]. Two intermediate nodes are added to the system, i.e. an application node and a function node. Gateway nodes store raw data generated by the sensors, whereas function nodes process the raw data. A service application connects directly with the function nodes.

This work uses NFV to service a node, rendering flexible and scalable micro-services. The proposed prototype is integrated – via cloud, edge and IoT devices – with a camera. Results of the experiment prove reduced processing time and improved resource sharing.

Different resource modeling approaches are summarized and compared in Table 1. One may see that efficient resource modeling approaches are very important to achieve portability, scalability and ease of maintenance. Semantic based modeling facilitates M2M communication and process automation. But due to multiple SDOs and their formats, interoperability is still a challenging task. Also, while building resource models, the level of abstraction should be selected legitimately. Otherwise, metadata management will become a tedious task. Resource modeling based on the attributes of resources may be used to build mathematical models for simulating and testing real world scenarios. As far as virtual resource modeling is concerned, it helps ease maintenance and offers better utilization of resources. Virtualization based SDN and NFV aid in handling heterogeneity and scalability.

2.4. Resource Discovery in IoT

The resource discovery module is responsible for identifying the resources required to satisfy application requests,

i.e. for device registration, configuration and ranking. Due to the large number of different IoT devices, the resource discovery module is expected to offer automatic discovery of resources, their properties and capabilities. Some of the existing technologies used here include P2P and distributed resource discovery, centralized architecture, CoAP-based, semantic based, and search engine based resource discovery. These are described below:

- **Distributed and P2P Discovery Services.** In this method, distributed devices are used to adopt a distributed hash table and are interlinked by a P2P overlay protocol to enable local and global service discovery. The method supports also multi attribute indexing, range queries and P2P routing. Therefore, the distributed architecture model effectively handles scalability and various issues related to IoT;
- **Centralized Architecture for Resource Discovery.** In this method, a central infrastructure is built for resource discovery. All devices should register to the centralized infrastructure before providing services. Client devices may discover and take advantage of a given service by accessing the central server;
- **CoAP-based Resource Discovery.** Here, the service discovery mechanism is adopted. CoAP servers expose a set of RESTful web services to clients. The interested CoAP clients request the server to provide service. Based on the request, a client may receive a list of resources and their attributes or metadata. But this method has certain drawbacks related to resource discovery lookup by remote devices. It also suffers from scalability and security issues;
- **Semantic-based Resource Discovery.** Resource discovery is based on semantic interpretation of resources. Standard web technologies are used for semantic representations of devices and their capabilities,

fostering interoperability, such as URI, HTTP, SensorML, JSON (JavaScript Object Notation) and efficient XML Interchange (EXI);

- **Search Engine-based Resource Discovery.** This model is similar to the Google search engine, but is not identical. The Google search engine is based on web documents, while an IoT search engine is based on resources, attributes and their properties. There are different types of IoT search engines, such as keyword, location, real-time, and hybrid search engines.

Paper [25] presents a search engine based resource discovery framework which is made up of three layers, namely proxy, discovery and service enablement layer. The proxy layer discovers the objects regardless of their technologies. It includes drivers and the protocol required to communicate with the devices. The discovery layer is responsible for storing metadata in the CoRE link format and for searching and ranking the resources based on user input. The service enablement layer provides the service to the user by RESTful web services. Interoperability between different description models of the sensor is a challenging task. Paper [26] describes a new framework based on keyword or geospatial searching approach. This framework supports API, SensorML and W3C JSON-LD description models. It will automatically detect the modeling language and parse its data. Then, this parsed data is converted into a standard format and stored in the database, along with location.

In [27], a centralized resource discovery framework is proposed, using web standards. In the proposed framework, resource categorization and indexing are done based on multiple domains. SDN based resource discovery in IoT networks is discussed in [28]. It suggests that traditional network architectures are not suitable for IoT resource discovery and routing path definition. To obtain a congestion free network, the author proposes an SDN based IoT net-

Table 2
Comparison of different resource discovery approaches in IoT

References	Method/approach	Centralized	Distributed	Framework/prototype/model proposed
[25]	Search engine based	•	○	New framework is proposed with proxy, discovery and service enablement modules.
[26]	Semantic-based indexing	•	○	The proposed framework parses metadata of multiple standards and creates indexes automatically.
[27]	Semantic-based	•	○	The proposed framework stores context-aware and geolocation information, along with resource information.
[28]	SDN-based	•	○	iFogSim simulation tool is used for efficient resource discovery based on SDN.
[29]	Agent-based	○	•	Multiple agents are installed in IoT devices, gateway and server.
[30]	Mathematical model	○	•	Effects of multiple context aware parameters are analyzed using a mathematical model.

work. Context information from network entities, computational devices and IoT devices is analyzed by means of a centralized SDN approach. The proposed model is simulated using the iFogSim tool which shows that SDN based IoT networks outperform traditional networks.

An autonomous multi-agent resource discovery mechanism for IoT is discussed in [29]. The author proposes a model with multiple agents running on the server, gateway and each IoT device. Initially, the server agent forwards the command to the gateway agent that forwards the command to agents of the particular devices. The proposed work assumed that agents are already preinstalled in each IoT device. This allows to avoid any commands that may be potentially missing during sensor sleep schedule, and IoT devices register their sleep pattern and type of service with the gateway agent.

The work presented in [30] analyses the different multi-objective decision making methods. The model is based on the centralized resource discovery model which uses the number of sensors and context information to ensure efficient results. The author claims that results obtained with the use of the simple additive weight method (SAW) offer better quality compared to the technique relying on the prioritization order, with that quality being similar to an ideal solution. In [31], the authors compare three commonly used discovery protocols, namely CoAP, MQTT and UPnP, based on memory footprint, CPU footprint, latency, downstream and upstream traffic. In this work, results of an experiment performed show that CoAP outperforms MQTT and UPnP, but at a higher memory overhead.

Table 2 summarizes different resource discovery approaches presented in recent publications.

2.5. Resource Estimation

In most IoT applications, user requests and data generation are spontaneous processes that pose a great challenge in the prediction of IoT resources. But for optimal system performance and efficient utilization, a certain level of estimation is not accurate enough. Generally, estimation of resources is performed based on historical records stored in the cloud. Fog may also aid in the estimation of resources, as they are deployed near to the sensing devices and they also have better knowledge of the location and higher context awareness.

IoT resource estimation depends on device type, device mobility, energy status, type of data generated or processed, communication method, security measures adopted and customer behavior. The extra pre-allocation of resources may lead to their underutilization. Improper allocation of resources will also directly affect system performance. Therefore, the resource management system should accurately predict the resources required for optimal performance of the system. In this regard, data analytical tools can be used for the prediction of data based on historical records stored in the cloud. The context information derived from edge and fog devices can also be used for accurate estimation of nearby resources. Data

mining algorithms, probabilistic methods, predictive algorithms, location- and context-aware resource approaches, etc. are used for this purpose.

Some research papers are devoted to IoT resource estimation. In [32], a mathematical model for determining customers' reliability and loyalty in order to ensure better utilization of resources is proposed. In this work, relinquish probability is used to prioritize loyal customers and for giving a fair chance to new ones. [33] proposes a resource estimation and allocation approach that is based on customer and device type. In the model, customers are categorized as new, absolutely new and existing. Paper [34] presents a resource allocation method based on customer type and loyalty. It uses crawled real trace and Amazon EC2 pricing. Results show that loyal customers are provided with better service and resources are allocated cautiously to other customers.

2.6. Resource Monitoring

IoT networks link a vast number of heterogeneous devices that support a wide range of applications. Real time applications, e.g. those used for patient health monitoring, flood alerting, fire alarms and for reporting loss of or delays in transferring information, may be a cause of a serious failure. Monitoring tools used in IoT should detect device failures, degraded performance, network delay, security attacks, etc. Algorithms are also needed for resource monitoring, (fault tolerance algorithms, security algorithms, agent-based algorithms, machine learning and so on).

In [35], the challenges related to IoT resource monitoring, e.g. identification of sensor malfunctions, inappropriate calibration, delays, packet loss, network failures, device energy status, security attacks and performance issues related to CPU and memory are discussed. [36] presents a new IoT monitoring tool for the agent-manager model. It is a lightweight solution compared with the SNMP based monitoring system. The agent devices follow commands forwarded by the manager device, i.e. a gateway.

Paper [37] outlines the IoT ecosystem challenges that are related to effective monitoring of resources. It also compares two standard network monitoring tools, namely big brother (BB) and Zenoss.

2.7. Resource Allocation

In a heterogeneous IoT system, achievement of a satisfactory level of service (SLA) related to end-user applications is a major challenge. For example, patient monitoring systems and city surveillance systems require completely different resource management strategies. When an application requests a service from IoT devices, multiple steps are performed before the resources are allocated. In short, client authentication requires specific resources, so the pool of available resources, access permission and utilization constraints need to be verified in the middleware framework. To solve complex resource allocation problems,

mathematical models are required that operate in the real world environment. In such models, the specific scenario is mapped to a standard optimization algorithm. The important categories of resource allocation algorithms include the following:

- **Multi-objective optimization algorithms.** These algorithms identify the best option from a set of feasible solutions. IoT resource allocation problems are often modeled as multi-objective optimization problems, since they involve a conflict of multiple criteria objectives. Most resource allocation problems try to apply the Pareto rule, where a change in any objective degrades at least one of the remaining objectives. Since these algorithms find exact solutions, they are computationally complex and consume more computational power;
- **Heuristic algorithms.** These algorithms are faster in finding an approximate solution instead of an accurate solution that requires a higher computation cost. They are also used to obtain near optimal solutions when other optimization algorithms fail to generate exact data. All these benefits are achieved by trading optimality, accuracy, precision and completeness of the solution. These algorithms lack in flexibility and are sensitive to key decision parameters. Even minor changes will drastically affect the end solution. Examples of such algorithms include, greedy and tabu search approaches;
- **Meta-heuristics algorithms.** These are higher level procedures relied upon to find a sufficiently good solution based on incomplete or imperfect information. These algorithms do not provide a globally optimal solution. Their examples include ant colony optimization, particle swarm optimization or genetic algorithms;
- **Game theory-based algorithms.** Game theory models rely on rational behaviors of two or more players and on their interdependence. Cooperative and non-cooperative games models are usually applied to solve resource allocation issues in IoT ecosystems by achieving the Nash equilibrium. The game theory model may not be suitable for a number of scenarios. For example, the game theory expects each player to be homogenous and to have similar capabilities, while IoT networks consist of highly heterogeneous devices. Also, it expects multiple interactions between the participating devices. Fulfillment of this condition is not feasible due to the large number of devices and the related energy constraints.

Among many papers focusing on resource allocation, it is worth mentioning [38] which discusses network lifetime and throughput for uplink communication in battery powered IoT networks. At first, the available channels are grouped and allocated to IoT users. Then, transmission

power allocation is optimized for each user group based on the Markov decision process (MDP). The efficient channel allocation algorithm (ECAA) proposed outperforms the random channel assignment approach by 90% and offers better time efficiency.

The problem of assigning a gateway to IoT devices is discussed in [39]. Each sensing device has to forward data either to the cloud or the server through the one or more gateways. Proper mapping of IoT devices and gateways allows to achieve good QoS. In the proposed work, the author modeled the gateway allocation problem as an NP-complete bipartite matching problem, where IoT devices have different data transmission requirements. Later, the work proposes a repetitive Ford-Fulkerson algorithm (RFF) for equal workload scenarios, and an approximation algorithm called greedy gateway algorithm (GGA) for unequal workload situations, minimizing the maximum traffic at the gateway.

Paper [40] presents a resource allocation problem experienced in the downlink connection of a fog-based IoT network. QoS parameters, such as delay, throughput and bit error rate are considered for ultra-reliable low latency communications (URLLC) and for enhanced mobile broadband (eMBB) service type applications. URLLC requires a high tolerable BER, whereas eMBB requires a high bandwidth which may comprise delay and BER. Centralized resource allocation will be NP-hard considering the heterogeneous nature and the number of devices involved. The integrated analytic hierarchy process (AHP) and the two-sides matching game approach ensure higher utility gains. AHP reduces complex QoS into traceable hierarchical subproblems, which helps in prioritizing QoS parameters, while the two-side matching game approach is initiated between fog network infrastructure and IoT devices. The service-to-interface assignment (SIA) problem that is extended onto multiple IoT devices and modeled as SIA-MID (service-to-interface assignment for multiple IoT devices) is presented in [41]. The classic SIA problem is designed for a single IoT resource, whereas SIA-MID is designed for more than one IoT device.

Bandwidth allocation between access points (APs) and the base station (BS) and modeled as a hierarchical two-stage game model is presented in [42]. At the first stage, bandwidth allocated to BS is distributed among APs using the asymptotic Shapley value method. Each AP acts as a game player forming an inter-AP game. In the second stage, bandwidth allocated to each AP is distributed across different classes of multimedia services using the relative utilitarian value approach. Different classes of traffic services will act as game players here. At the end of each unit's time period, the relative utilitarian value is recalculated and bandwidth is redistributed across APs. When compared with the game-theoretic hierarchical resource allocation (GHRA) scheme and the auction-based hierarchical resource allocation (AHRA) model, the algorithm shows better throughput, service rate and minimal resource utilization.

[43] proposes a new resource allocation mechanism called search economics-based IoT resource allocation (SEIRA) and using meta-heuristic and data clustering algorithms. In the data clustering algorithm, k-means is used to generate improvised initial values that are used in search economics (SE). SEIRA uses hybrid encoding, modified transit and a local search operator for better results. SEIRA and SEIRA-KD (SEIRA with k-means and dynamic local search) yield better results and are faster compared with GA and SA methods. An optimal resource allocation strategy for the three-tier IoT architecture is proposed in [44]. The three-tier architecture includes data service operator (DSO), data service subscribe (DSS) and fog nodes (FN), where each entity acts autonomously and observes the behaviors of each other for better decision making. The proposed algorithm has four stages, including resource purchasing problem for DSSs, with DSOs modeled as the Stackelberg game, pricing problem for the DSOs, DSO and FN pairing problem as many-to-many matching games, and under same DSOs pairing problem between FN and DSS as many to many matching problems. Simulation results show that DSOs, FNs, and DSSs achieve higher utility rates autonomously.

A distributed consensus-based algorithm in which virtual objects (VOs) negotiate with other elements in order to optimize resource allocation by distributing the workload is presented in [45]. It extends the VOs model to incorporate quality of information (QoI) for decision making. The result shows an improved node lifetime and a faster convergence time for task distribution. The fog radio access network (F-RAN) resource allocation problem is discussed in [46]. Fog RAN reduces latency – an important parameter in IoT applications. The proposed algorithm is modeled as MDP and is further reinforcement by learning to obtain the optimum solution. Simulations of the proposed algorithm outperforms the fixed threshold algorithm.

Paper [47] proposes a middleware-based approach used to distribute the task among different IoT devices, so that resources are adequately shared. The proposed middleware has two layers, i.e. semantic layer for interoperability of nodes, as well as resource allocation and management layer (RAML) for task group management. The distributed protocol is proposed based on the consensus algorithm which converges the optimal allocation of resources offering, in most cases, an error rate of less than 5%. The proposed model is tested in Matlab, using the Telit development board. Data transmission between gateways is presented in [48]. Composite services are executed at the gateway by using one or more IoT resources that are allocated to at least one gateway. To reduce the volume of data exchanged between gateways while executing composite services, network topology is transformed into the degree-constrained minimum spanning tree (DCMST). A modified version of DCMST is used to identify the optimal solution by applying the genetic algorithm. Simulations of the proposed algorithm show a 97% success rate, on average, while finding a near optimal solution. Allocation of heterogeneous

IoT resources is modeled as the Stackelberg game model in [49]. The network operator is modeled as the leader and the mobile terminal as a follower. The network operator sells the resources to the mobile terminal based on dynamic pricing. A distributed iterative algorithm is proposed which allows the entire network to reach the Nash equilibrium as a part of the Stackelberg game model. A simulation result provides an optimal pricing strategy.

The work presented in [50] proposes an adaptive resource allocation algorithm (AMSRS) for IoT devices with constrained bandwidth. When performing sensing and actuating operations, network bandwidth is dynamically assigned based on the signals' frequency domain characteristics. Compared with the static algorithm, AMSRS reduces accumulated and maximum errors by 60%, while path tracking the UGV. A novel resource management framework is proposed based on the device cloud approach [51]. In the proposed framework, IoT resources are considered as part of a cloud resource pool. The proposed work considers IoT end devices as data end points (DEP), the gateway as a data integration point (DEP), and the communication infrastructure as a machine communication network (MCN). The core components of the proposed framework include device directory which provides a data model for the devices and their life cycle. The user directory module is designed for identifying and controlling access to management services. The management module is developed for performing the resource allocation process. The proposed framework is implemented at the gateway or an aggregating node. Article [52] discusses the challenges that need to be tackled in the device cloud approach, such as runtime device integration, interoperability, data models, nomenclature, and the service execution environment.

Paper [53] offloads the task from the local wireless device to a fog network. The different parameters to be managed at the device and network level are modeled using mixed integer linear programming (MINLP). The hybrid genetic simulated annealing algorithm is applied to solve the proposed issue, resulting in lower power consumption by local devices, faster convergence rates and minimum latency. An energy efficient resource optimization algorithm for cognitive IoT (CIoT) is proposed in [54]. Since CIoT consumes more energy for sensing operations in order to achieve better spectrum efficiency, the proposed algorithm harvests the primary user's RF energy to increase network lifetime. The number of nodes, sensing time and transmission power are modeled as a non-convex optimization problem. The proposed work suggests a node alternative algorithm which uses Langrange optimization for power adjustments and Dinkelbach's optimization for node energy and network lifetime corrections.

Paper [55] presents an application-aware workload distribution approach focusing on edge devices. Resources of edge devices are dynamically adjusted to reduce response time. A heuristic algorithm is applied to distribute user requests to different edge devices. Research presented in [56] proposes a framework for multiband spectrum sensing,

Table 3
Comparison of different resource allocation approaches

References	Mathematical model (MM)/ prototype framework (PFW)	Resource allocation infrastructure	Scalability	Heterogeneity	Load balancing	Real time	QoS	Fault tolerance	Mobility	Performance
[38]	MM – Markov decision process	- gateway with LoRa - IoT device	○	○	○	●	○	○	●	- network lifespan - throughput
[39]	MM – greedy method (heuristic method)	- IoT device - gateway	●	○	●	○	○	○	○	traffic minimization
[40]	MM – game theory	- fog devices - IoT device - gateway	●	●	○	○	●	○	○	- throughput - utilization of fog devices
[41]	MM – metaheuristic method	- IoT device - gateway	●	●	●	○	○	○	○	faster convergence
[42]	MM – game theory	- access point - IoT device - gateway	●	○	●	●	●	○	●	- throughput - service rate - better resources utilization
[43]	MM – metaheuristic method	- IoT device - gateway	●	○	●	○	○	○	○	faster solution convergence rate
[44]	MM – game theory	- IoT device - fog network - data service provider network	●	○	●	○	○	●	○	higher utilization at fog and IoT devices
[45]	Framework based on distributed consensus-based algorithm	- cloud - IoT devices	●	○	●	○	○	○	○	- node lifetime - faster convergence
[46]	MM – Markov decision process	fog node	●	●	○	●	○	○	○	- latency - utilization
[47]	Middleware framework prototype	- cloud - IoT device	●	○	●	○	○	○	○	- faster convergence - less error rate
[48]	MM – metaheuristic method	- IoT - gateway	○	●	●	○	○	○	○	- faster convergence - minimize traffic
[49]	MM – game theory	- network operator infrastructure - IoT device	●	●	○	○	○	○	○	utilization
[50]	MM – multi objective optimization	IoT device	○	○	○	●	●	○	●	error rate reduction
[53]	MM – metaheuristic method	- fog devices - gateway - IoT device	○	●	●	○	○	○	○	- node energy - latency
[54]	MM – heuristic method	- gateway - IoT device	○	○	●	○	●	○	○	- network lifetime - node energy
[55]	MM – heuristic method	- fog devices - gateway - IoT device	●	●	●	○	○	○	○	latency
[56]	MM – heuristic method and game theory	- gateway - IoT device	○	●	○	○	●	○	○	faster convergence
[57]	MM – game theory	- gateway - IoT device	●	●	○	○	●	○	○	- node energy - packet to data ratio
[58]	MM – metaheuristic method	- gateway - IoT device	●	○	●	○	○	○	○	- service rate - node energy
[59]	Block chain based framework	- gateway - IoT device	●	●	○	○	○	○	○	computation cost
[60]	MM – heuristic method	- fog devices - gateway - IoT device	●	●	●	○	○	○	○	- service rate - utilization
[61]	MM – heuristic method and game theory	- gateway - IoT device	●	○	○	●	●	○	○	- data rate - node energy
[62]	MM – heuristic method	- gateway - IoT device	●	○	○	○	●	○	○	- data rate - node energy
[63]	MM – metaheuristic method	- fog devices - gateway - IoT device	●	○	●	○	○	○	○	- energy - utilization

using the branch and bound algorithm. It provides a cross-layer configuration scheme for channel allocation based on the potential game. The proposed algorithm helps achieve dynamic resource allocation and also improves QoS parameters, such as energy consumption, throughput, delay, reliability and cost. In [57], the author provides energy efficient and cost-effective solutions for threat models using Pareto and Stackelberg leadership games. In the suggested model, the packet-to-data ratio (PDR), energy and the cost of installation are improved. The proposed model is tested using the TOSSIM simulator.

Article [58] proposes a task offloading and resource allocation model for mobile edge computing in an unmanned aerial vehicle (UAV). The proposed application involves resource intensive tasks that need to be performed with minimum delay, which is an additional constraint. In the proposed work, UAV position, uplink, downlink and computation resources are modeled as an optimization problem. Simulation results show a reduced service delay and lower energy consumption by IoT devices. In [59], a Blockchain based resource allocation approach is proposed, relying on edge computing. Edge computing devices act as a service provider for mobile devices which are willing to run their miners. In the model presented, a new hash-based power function is defined and an auction is conducted to reduce computational cost and to increase social welfare. [60] presents a method for balancing the load of fog devices. To reduce the load, services are migrated dynamically. The proposed model is tested on CloudSim, showing improved resource utilization, service rate and a reduced number of service nodes required. Paper [61] improves the data rate and energy utilization by using the heuristic greedy method and the cooperative game theory in the IoT network. The greedy method is used to find the optimal local solution, whereas interference between radio signals generated by IoT devices is reduced by means of the game theory.

For adaptive carrier spectrum allocation and joint spreading time code generation, the joint step extreme recursion (JSER) algorithm is proposed in [62]. This anti-jamming self-adaptive algorithm reduces power consumption and improves the signal to interference plus noise ratio (SINR). The computational and cache resource management problem is modeled as a joint optimization problem in [63]. The SDN-based information-centric model takes the allocation decision based on the Q-deep learning artificial intelligence algorithm. Dynamically short-term and long-term rewards are provided for energy and computational cost expenditures.

Table 3 presents a comparison of the different resource allocation approaches. To summarize, the majority of research is based on considering resource allocation problems, such as resource optimization and middleware framework. In the first approach, many researchers have developed mathematical models using the graph theory, relying on operation research applying different meta-heuristic approaches, such as ant colony optimization, particle swarm optimization and

Table 4

Major metrics used in resource allocation approaches

Metrics	References
Network lifetime	[38], [53]
Throughput	[38], [40], [42], [59]
Data traffic	[39], [48]
Service rate	[42], [56], [58]
Latency	[46], [51], [53]
Error rate	[47], [49]
Device utilization	[40], [42], [44], [45], [46], [49], [58], [61]
Computation cost	[41], [43], [47], [48], [54], [57]
Service rate	[42]
Node lifetime	[45], [51], [52], [55], [56], [59], [60], [61]

Table 5

Architecture types used in resource allocation

Metrics	References
Two-tier architecture	[38], [39], [41], [42], [43], [45], [48], [49], [50], [52], [54], [55], [56], [57], [59], [60]
Three-tier architecture	[40], [44], [46], [47], [51], [53], [58], [61]

genetic algorithm. Many authors also use different game theory approaches, such as cooperative game, evolutionary game and Stackelberg game, where interactions between the data service provider, access point, fog and IoT devices are modeled as conflicting games. In the other approach, most of the work is based on developing a middleware prototype for resource management purposes. Only a few works focus on testing resource allocation strategies using the standard framework. But testing with the use of the real framework is really important to understand the behavior of the proposed algorithm and its assumptions in a real world scenario. In Table 4, the important resource allocation metrics are presented in terms of scalability, reduced communication and computation cost and load balancing aiming to improve node and network lifetime. Edge and fog computing are promising solutions for highly computationally-intensive operations that delay sensitive applications. Table 5 compares the number of resource allocation schemes proposed for 3-tier and 2-tier architectures.

3. IoT Resource Management Module in FIWARE

To manage diverse technologies and components of IoT, generic enablers (GEs) or abstract interfaces have been developed. For example, the European ICT and standardization committee proposes an open-source middleware platform known as FIWARE [64]. Another popular open-source project – IoTivity – proposed by the Open Connec-

tivity Foundation (OCF) [65], also outlines similar responsibilities of the resource management module. Both architectures incorporate resource identification and addressing schemes, mapping of device ID to a specific device, dynamic linking of resources, as well as mechanisms for controlling access to resources and information in the resource management module. They also include device management functions, such as activation, deactivation, software update or remote monitoring. The FIWARE architecture proposes four GEs, such as communication, resource management, data handling and process automation. The resource management process is shown in Fig. 8.

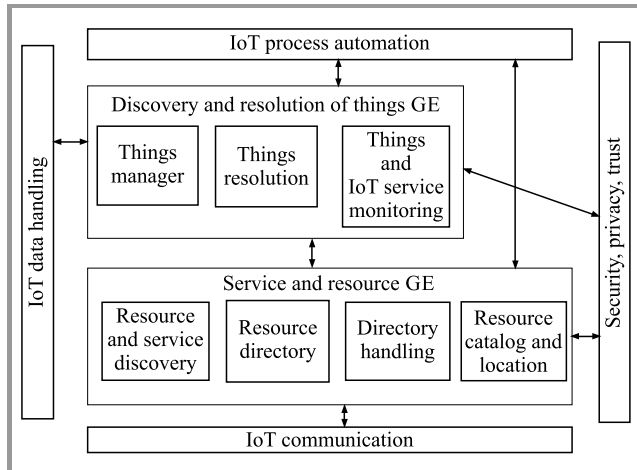


Fig. 8. Resource management using a generic enabler (GE).

This abstraction layer provides unified service and operation support for the management of IoT resources. A resource management GE is used and is responsible for identifying, registering, unregistering, discovering, activating, utilizing and configuring resources. It consists of such sub-layers as discovery, resolution of things, as well as services and resources interaction. These are described as follows:

- **Discovery and resolution of things generic enablers.** They help discover services and resources based on the device ID.
- **Services and resources interaction generic enablers.** These provide such services as registration, deregistration, status retrieval and updates performed by sub-modules. Before providing a given service, the enabler interacts with the IoT process automation module for context and mobility information. It also interacts with the security and privacy module to verify access permission for a given request.

4. Challenges and Research Issues

Each year a great number of smart objects is integrated with the IoT ecosystem. In the view of the scale and heterogeneity of the phenomenon, there is a need for dynamic, scalable and robust solutions allowing to manage IoT resources. After an extensive survey of existing solutions,

we have identified the following major issues related to resource management and requiring further research to be performed:

- Interoperability of IoT resources used on different platforms – there is no common description for resources, their capabilities and format of data exchanged between different SDOs. Interoperability across multiple SDOs simplifies the resource management process across multiple domains;
- Prioritization of discovered resources helps the resource management module select the appropriate resource;
- Security – standard security mechanism for resource discovery;
- Load balancing – for optimal utilization of resources, a distributed resource management module is required that is capable of relying on the benefits offered by fog and edge computing schemes;
- Dynamic task distribution mechanisms – based on the capabilities of the participating devices. There is great room for novel dynamic task management modules based on the resources available within the system;
- Resource estimation – to manage the resources effectively;
- Predictive mechanisms/algorithms – used in resource allocation;
- Resource scheduling and allocation – for real time IoT applications;
- Improved pricing model – for resources used in cloud, fog, edge and for IoT;
- Fault tolerance – to provide uninterrupted services to the user during node failure. Most of the proposed analytical models fail to accurately map actual resource allocation problems experienced in IoT. Also, many assumptions made while modeling the problem fail to address the major challenges, just as the majority of research does not discuss ways to integrate the local optimal solution identified with global solutions of resource allocation-related problems.

No unified standard exists accepted across the entire market. Individual SDOs and industry alliances have defined their own formats. This prevents end users from being able to take advantage of improved services. To ensure cooperation between SDOs, SSOs and industry alliances and to identify gaps between them, special groups – such as the ESTI specialist task group [66], the alliance of Internet of Things innovations (AIOTTI) [67] or CREATE-IOT [68] have been formed. Even though work towards standardization is already in progress, adaptation of the results on the

market is still very low. To address these issues, there is an urgent need for research to develop a dynamic and efficient broker capable of understanding and mapping multiple standard resources. Also, there is a need to provide software developers with fine-grained descriptions in order to avoid ambiguity caused by abstract information from SDOs.

The management of IoT resources used in each layer of the environment poses a number of challenges. Most work is performed to develop mathematical resource allocation models. But these proposed models are tested only with the use of simulation tools or software developed in-house. So, there is an immediate need for framework or test bed based verification of algorithms.

Semantic web-based modeling, although enhancing interoperability, also suffers from incompatibility between multiple SDOs, SSOs and industrial alliances. Semantic-based technologies experience a range of other issues, such as the need for lightweight semantic technology, improved security mechanism for registration and resource discovery, large indexing and querying capacity, dynamic conversion between standards, efficient streaming models and open standards or interfaces for application developers.

5. Conclusion

In this paper, heterogeneity of IoT devices, distributed architectures of IoT ecosystems, types of resources, challenges in managing the heterogeneous resources, phases of resource management in IoT, guidelines provided by SDOs for interoperability in managing resources and potential design options available in each resource management phase (i.e. semantic- or virtualization-based modeling, centralized or distributed resource discovery) are discussed.

The key outcome of this study indicates that most algorithms discussed in the resource allocation section are mathematical or attribute-based. Assumptions made and results obtained with the use of these algorithms are hard to verify in real frameworks or testbeds. Our immediate research will focus on proposing a resource allocation algorithm considering the distributed architecture of the IoT ecosystem. The proposed algorithm will be verified using the existing IoT framework. Furthermore, open issues related to resource management schemes, as discussed in this paper, are expected to motivate researchers to conduct further investigations in this particular area.

References

- [1] A. Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications", *IEEE Communi. Surv. Tutor.*, vol. 17, no. 4, pp. 2347–2376, 2015 (DOI: 10.1109/COMST.2015.2444095).
- [2] M. Burhan, R. A. Rehman, B. Khan, and B. Kim, "IoT elements, layered architectures and security issues: a comprehensive survey", *J. of Sensors*, vol. 18, no. 9, pp. 2796–831, 2018 (DOI: 10.3390/s18092796).
- [3] "IoT Reference Model" [Online]. Available: <https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=y.2060>
- [4] F. Delicato, P. Pires, and T. Batista, *Resource Management for Internet of Things*. Springer, 2017 (ISBN: 9783319542478).
- [5] Z. Ghanbari, N. J. Navimipour, M. Hosseinzadeh, and A. Darwesh, "Resource allocation mechanisms and approaches on the Internet of Things", *J. of Cluster Comput.*, vol. 22, no. 4, pp. 1253–1282, 2019 (DOI: 0.1007/s10586-019-02910-8).
- [6] A. Botta, W. Donato, V. Persico, and A. Pescapé, "Integration of cloud computing and Internet of Things", *J. of Fut. Gener. Com. Syst.*, vol. 56, pp. 684–700, 2016 (DOI: 10.1016/j.future.2015.09.021).
- [7] "ITU Common Functional and Non Functional requirement of IoT" [Online]. Available: <https://www.itu.int/rec/T-REC-Y.2066>
- [8] M. Glinz, "On non-functional requirements", in *Proc. of 15th IEEE Int. Requir. Engin. Conf.*, Delhi, India, 2007, pp. 21–22 (DOI: 10.1109/RE.2007.45).
- [9] A. N. Lam and O. Haugen, "Applying semantics into service-oriented IoT framework", in *Proc. of IEEE 17th Int. Conf. on Indust. Inform. INDIN 2019*, Helsinki, Finland, 2019, pp. 206–213 (DOI: 10.1109/INDIN41052.2019.8972295).
- [10] A. Sheth, C. Henson, and S. Sahoo, "Semantic sensor web", *IEEE J. on Internet Comput.*, vol. 12, no. 4, pp. 1–10, 2008 (DOI: 10.1109/MIC.2008.87).
- [11] P. Barnaghi, M. Presser, and K. Moessner, "Published linked sensor data", in *Proc. of Int. Symp. on Collabor. Technol. and Syst.*, pp. 1–16, 2010, in ISWC 2010, 2010-11-07–2010-11-11, Shanghai, China [Online]. Available: <http://epubs.surrey.ac.uk/470673/1/sense2web.pdf>
- [12] I. Alam *et al.*, "IoT virtualization: A survey of software definition and function virtualization techniques for Internet of Things", pp. 1–30, 2019 [Online]. Available: <https://arxiv.org/pdf/1902.10910.pdf>
- [13] B. I. Ismail, "Evaluation of docker as edge computing platform", in *Proc. of IEEE Int. Conf. on Open Syst.*, Bandar Melaka, Malaysia, 2015, pp. 130–135 (DOI: 10.1109/ICOS.2015.7377291).
- [14] J. G. Ko *et al.*, "Sensor virtualization module: virtualizing IoT devices on mobile smart phones for effective sensor data management", *J. of Distrib. Sensor Netw.*, vol. 11, no. 3, pp. 1–10, 2015 (DOI: 10.1155/2015/730762).
- [15] B. Billet and V. Issarny, "From task graphs to concrete actions: A new task mapping algorithm for the future Internet of Things", in *Proc. of IEEE 11th Int. Conf. on Mob. ad hoc and Sensor Syst.*, Philadelphia, PA, USA, 2014, pp. 470–478 (DOI: 10.1109/MASS.2014.20).
- [16] V. Angelakis, I. Avgouleas, N. Pappas, E. Fitzgerald, and D. Yuan, "Allocation of Heterogeneous resources of an IoT device to flexible services in IoT", *J. on IEEE Internet of Things*, vol. 3, no. 5, pp. 691–700, 2016 (DOI: 10.1109/IIOT.2016.2535163).
- [17] M. Sharief, O. Kingston, S. Hossam, and O. Kingston, "Resource re-use in wireless sensor networks: realizing a synergetic Internet of Things", *J. of Commun.*, vol. 7, no. 7, pp. 484–493, 2012 (DOI: 10.4304/jcm.7.7.484-493).
- [18] P. Banerjee *et al.*, "Everything as a service: powering the new information economy", *J. of Comp.*, vol. 44, no. 3, pp. 36–43, 2011 (DOI: 10.1109/MC.2011.67).
- [19] W. Wang, S. De, G. Cassar, and K. Moessner, "Knowledge representation in the Internet of Things: semantic modeling and its applications", *Automatika J. for Control, Measur., Electron., Comput.*, vol. 54, no. 4, pp. 388–400, 2013 (DOI: 10.7305/automatika.54-4.414).
- [20] J. Soldatos *et al.*, "OpenIoT: open source Internet-of-Things in the cloud", in *Interoperability and Open-Source Solutions for the Internet of Things*, LNCS, vol. 9001, pp. 13–25. Springer, 2015 (DOI: 10.1007/978-3-319-16546-2_3).
- [21] M. Compton *et al.*, "The SSN Ontology of the W3C Semantic Sensor Network", *J. of Web Semant.*, vol. 17, pp. 25–32, 2012 (DOI: 10.1016/j.websem.2012.05.003).
- [22] S. Nagowah and B. A. Rahimbux, "An overview of semantic interoperability ontologies and frameworks for IoT", in *Proc. of IEEE 6th Int. Conf. on Enterpr. Syst. ES 2018*, Limassol, Cyprus, 2018, pp. 470–478 (DOI: 10.1109/ES.2018.00020).

- [23] A. Palade, C. Cabrera, G. White, M. A., Razzaque, and S. Clarke, "Middleware for Internet of Things: A quantitative evaluation in small scale", in *Proc. of IEEE 18th Int. Symp. on a World of Wirel., Mob. and Multim. Netw. WoWMoM 2017*, Macau, China, 2017, pp. 1–6 (DOI: 10.1109/WoWMoM.2017.7974340).
- [24] K. Ogawa, K. Kanai, K. Nakamura, H. Kanemitsu, J. Katto, and H. Nakazato, "IoT device virtualization for efficient resource utilization in smart city IoT platform", in *Proc. of IEEE Int. Conf. on Perv. Comput. and Commun. Worksh. PerCom Workshops 2019*, Kyoto, Japan, 2019, pp. 419–422 (DOI: 10.1109/PERCOMW.2019.8730806).
- [25] S. K. Datta, R. P. Costa, and C. Bonnet, "Resource discovery in Internet of Things: Current trends and future standardization aspects", in *Proc. of IEEE 2nd World Forum on Internet of Things WF-IoT 2015*, Milan, Italy, 2015, pp. 542–547 (DOI: 10.1109/WF-IoT.2015.7389112).
- [26] N. Hussain, T. Anees, and AzeemUllah, "Development of a novel approach to search resources in IoT", *Int. J. of Adv. Comp. Sci. and Appl.*, vol. 9, no. 9, pp. 385–398, 2018 (DOI: 10.14569/IJACSA.2018.090949).
- [27] C. Perera and A. V. Vasilakos, "A knowledge-based resource discovery for Internet of Things", *J. of Knowledge-Based Syst.*, vol. 109, pp. 122–136, 2016 (DOI: 10.1016/j.knosys.2016.06.030).
- [28] M. Afrin and R. Mahmud, "Software defined network-based scalable resource discovery for Internet of Things", *J. of EAI Endorsed Trans. on Scalable Inform. Syst.*, vol. 17, no. 14, pp. 1–6, 2017 [Online]. Available: <https://eudl.eu/pdf/10.4108/eai.25-9-2017.153149>
- [29] P. Krivic, P. Skocir, and M. Kusek, "Agent-based approach for energy-efficient IoT services discovery and management", in *Agents and Multi-Agent Systems: Technologies and Applications 2018 Proceedings of the 12th International Conference on Agents and Multi-Agent Systems: Technologies and Applications (KES-AMSTA-18)*. Springer, 2018, pp. 57–66 (DOI: 10.1007/978-3-319-92031-3_6).
- [30] L. Nunes, J. Estrella, C. Perera, S. Reiff-Marganiec, and A. Delbem, "Multi-criteria IoT resource discovery: a comparative analysis", *J. on Software-Practice and Exper.*, vol. 47, no. 10, pp. 1325–1341, 2016 (DOI: 10.1002/spe.2469).
- [31] K. Khalil, K. Elgazzar, and M. Bayoumi, "A comparative analysis on resource discovery protocols for the Internet of Things", in *Proc. of IEEE Global Commun. Conf. GLOBECOM-2018*, Abu Dhabi, United Arab Emirates, 2018 (DOI: 10.1109/GLOCOM.2018.8647553).
- [32] M. Aazam, M. St-Hilaire, C. Lung, I. Lambadaris, and E. Huh, "IoT resource estimation challenges and modeling in fog", in *Fog Computing in the Internet of Things*. Springer, 2018, pp. 17–31 (DOI: 10.1007/978-3-319-57639-8_2).
- [33] M. Aazam and E. N. Huh, "Fog computing micro data center based dynamic resource estimation and pricing model for IoT", in *Proc. of the 29th IEEE Int. Conf. on Adv. Inform. Netw. and Appl.*, Gwangju, South Korea, 2015, pp. 687–694 (DOI: 10.1109/AINA.2015.254).
- [34] M. Aazam and E. Huh, "Resource management in media Cloud of Things", in *Proc. of IEEE Int. Conf. on Parall. Process. Worksh.*, Minneapolis, MN, USA, 2014, pp. 361–367 (DOI: 10.1109/ICPPW.2014.54).
- [35] M. Aazam and E. Huh, "Dynamic resource provisioning through fog micro data center", in *Proc. of the IEEE Int. Conf. on Perv. Comput. Commun. Worksh. PerCom Workshops 2015*, St. Louis, MO, USA, 2015, pp. 105–110 (DOI: 10.1109/PERCOMW.2015.7134002).
- [36] M. Aazam, M. St-Hilaire, C. Lung, and I. Lambadaris, "PRE-Fog: IoT trace based probabilistic resource estimation at fog", in *Proc. of the 13th IEEE Ann. Consumer Commun. and Netw. Conf. CCNC 2016*, Las Vegas, NV, USA, 2016, pp. 12–17 (DOI: 10.1109/CNC.2016.7444724).
- [37] E. Solaiman, R. Ranjan, P. Jayaraman, and K. Mitra, "Monitoring Internet of Things application ecosystems for failure", *J. on IEEE IT Profess.*, vol. 18, no. 5, pp. 1–4, 2016 (DOI: 10.1109/MITP.2016.90).
- [38] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered IoT networks", *IEEE Internet of Things J.*, vol. 6, no. 3, pp. 4935–4945, 2019 (DOI: 10.1109/JIOT.2019.2895417).
- [39] S. Manakkadu and S. Dutta, "On efficient resource allocation in the Internet of Things environment", in *Proc. of 8th Int. Conf. on the Internet of Things*, Santa Barbara, CA, USA, 2018, no. 22, pp. 1–5 (DOI: 10.1145/3277593.3277623).
- [40] S. F. Abedin *et al.*, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network", *J. on IEEE Trans. on Commun.*, vol. 67, no. 1, pp. 489–502, 2019 (DOI: 10.1109/TCOMM.2018.2870888).
- [41] C. Tsai and S. Liu, "An effective IoT service-to-interface assignment algorithm via search economics", *J. on IEEE Internet of Things*, vol. 5, no. 3, pp. 1708–1718, 2018 (DOI: 10.1109/JIOT.2018.2796310).
- [42] S. Kim, "Asymptotic shapley value based resource allocation scheme for IoT services", *J. on Comp. Netw.*, vol. 100, pp. 55–63, 2016 (DOI: 10.1016/j.comnet.2016.02.021).
- [43] C. Tsai, "SEIRA: An effective algorithm for IoT resource allocation problem", *J. on Comp. Commun.*, vol. 119, pp. 156–166, 2017 (DOI: 10.1016/j.comcom.2017.10.006).
- [44] H. Zhang, Y. Xiao, S. Bu, D. Niyato, F. R. Yu, and Z. Han, "Computing resource allocation in three-tier IoT fog networks: A joint optimization approach combining Stackelberg game and matching", *J. on IEEE Internet of Things*, vol. 4, no. 5, pp. 1204–1215, 2017 (DOI: 10.1109/JIOT.2017.2688925).
- [45] V. Pilloni and L. Atzori, "Consensus-based resource allocation among objects in the Internet of Things", *J. on Annals of Telecommun.*, vol. 72, pp. 415–429, 2017 (DOI: 10.1007/s12243-017-0583-6).
- [46] A. Nassar and Y. Yilmaz, "Resource allocation in fog RAN for heterogeneous IoT environments based on reinforcement learning", in *Proc. of IEEE Int. Conf. on Commun. ICC 2019*, Shanghai, China, 2019, pp. 1–6 (DOI: 10.1109/ICC.2019.8761626).
- [47] G. Colistra, V. Pilloni, and L. Atzori, "The problem of task allocation in the Internet of Things and the consensus-based approach", *J. of Comp. Netw.*, vol. 73, pp. 98–111, 2014 (DOI: 10.1016/j.comnet.2014.07.011).
- [48] M. Kim and I. Ko, "An efficient resource allocation approach based on a genetic algorithm for composite services in IoT environments", in *Proc. of IEEE Int. Conf. on Web Services*, New York, NY, USA, 2015, pp. 543–550 (DOI: 10.1109/ICWS.2015.78).
- [49] H. Lan, H. Song, H. Liu, and G. Y. Zhang, "Heterogeneous-oriented resource allocation method in Internet of Things", *J. of Appl. Mechan. and Mater.*, vol. 427, pp. 2791–2794, 2013 (DOI: 10.4028/www.scientific.net/AMM.427-429.2791).
- [50] L. Zheng, L. Kaihua, S. Yuting, and M. Yongtao, "Adaptive resource allocation algorithm for Internet of Things with bandwidth constraint", *J. on Trans. of Tianjin Univer.*, vol. 18, pp. 253–258, 2012 (DOI: 10.1007/s12209-012-1873-8).
- [51] A. Kliem and O. Kao, "The Internet of Things resource management challenge", in *Proc. of IEEE Int. Conf. on Data Sci. and Data Intensive Syst.*, Sydney, Australia, 2015, pp. 483–490 (DOI: 10.1109/DS-DIS.2015.21).
- [52] T. Renner, A. Kliem, and O. Kao, "The device cloud – applying cloud computing concepts to the Internet of Things", in *Proc. of IEEE 11th Int. Conf. on Ubiquit. Intell. and Comput., IEEE 11th Int. Conf. on Autonom. and Trust. Comput. and IEEE 14th Int. Conf. on Scal. Compu. and Commun. and Its Assoc. Worksh.*, Bali, Indonesia, 2014, pp. 396–401 (DOI: 10.1109/UIC-ATC-ScalCom.2014.106).
- [53] Q. Wang, and S. Chen, "Latency-minimum offloading decision and resource allocation for fog-enabled Internet of Things networks", *J. of Trans. on Emerg. Telecommun. Technol.*, 2020 [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3880> (DOI: 10.1002/ett.3880).
- [54] X. Liu, Y. Li, X. Zhang, W. Lu, and M. Xiong, "Energy-efficient resource optimization in green cognitive Internet of Things", *J. of Mob. and Netw. Appl.*, 2020 (DOI: 10.1007/s11036-020-01510-w).
- [55] Q. Fan and N. Ansari, "Application aware workload allocation for edge computing-based IoT", *J. on IEEE Internet of Things*, vol. 5, no. 3, pp. 2146–2153, 2018 (DOI: 10.1109/JIOT.2018.2826006).

- [56] W. Ejaz and M. Ibnkahla, "Multiband spectrum sensing and resource allocation for IoT in cognitive 5G networks", *J. of IEEE Internet of Things*, vol. 5, no. 1, pp. 150–163, 2018 (DOI: 10.1109/JIOT.2017.2775959).
- [57] A. Rullo, D. Midi, E. Serra, and E. Bertino, "Pareto optimal security resource allocation for Internet of Things", *J. on ACM Trans. on Priv. and Secur.*, vol. 20, no. 4, pp. 1–30, 2017 (DOI: 10.1145/3139293).
- [58] Z. Yu, Y. Gong, S. Gong, and Y. Guo, "Joint task offloading and resource allocation in UAV-enabled mobile edge computing", *J. on IEEE Internet of Things*, vol. 7, no. 4, pp. 3147–3159, 2020 (DOI: 10.1109/JIOT.2020.2965898).
- [59] Y. Jiao, P. Wang, D. Niyato, and Z. Xiong, "Social welfare maximization auction in edge computing resource allocation for mobile blockchain", in *Proc. of IEEE Int. Conf. on Commun. ICC 2018*, Kansas City, MO, USA, 2018 (DOI: 10.1109/ICC.2018.8422632).
- [60] X. Xu *et al.*, "Dynamic resource allocation for load balancing in fog environment", *J. on Wirel. Commun. and Mob. Comput.*, vol. 2, 2018 (DOI: 10.1155/2018/6421607).
- [61] H. Malik *et al.*, "Radio resource management scheme in NB-IoT systems", *J. of IEEE Access*, vol. 6, pp. 15051–15064, 2018 (DOI: 10.1109/ACCESS.2018.2812299).
- [62] Z. Dou, G. Si, Y. Lin, and M. Wang, "An adaptive resource allocation model with anti-jamming in IoT network", *J. of IEEE Access*, vol. 7, pp. 93250–93258, 2019 (DOI: 10.1109/ACCESS.2019.2903207).
- [63] F. Xu, F. Yang, S. Bao, and C. Zhao, "DQN inspired joint computing and caching resource allocation approach for software defined information-centric Internet of Things network", *J. of IEEE Access*, vol. 7, pp. 61987–61996, 2019 (DOI: 10.1109/ACCESS.2019.2916178).
- [64] FIWARE Open source Platform [Online]. Available: <https://www.fiware.org/>
- [65] IoTivity Open source Platform [Online]. Available: <https://iotivity.org/>
- [66] European Special Task Force 505 for IoT standards landscaping and IoT European Large Scale Pilot gap analysis [Online]. Available: <https://portal.etsi.org/STF/STFs/STFHomePages/STF505>
- [67] Standardization Gaps by The Alliance for Internet of Things Innovation [Online]. Available: <https://aioti.eu/aioti-wg03-reports-on-iot-standards/>
- [68] European Large Scale Pilot Programme [Online]. Available: <https://european-iot-pilots.eu/>



Lokesh B. Bhajantri received his Ph.D. degree in Computer Science and Engineering from the Visveswaraiiah Technological University (VTU), Belgaum, Karnataka, in 2015. He has been working, for 15 years now, as an Associate Professor at the Department of Information Science and Engineering,

Basaveshwar Engineering College, Bagalkot, India. His areas of interest include distributed/wireless sensor networks, cognitive Internet of Things, mobile computing and communications, networking protocols, genetic algorithms, applications of agents, and real time systems. He is a member of the Board of Computer Science, Engineering and Information Science Studies. He is also a member of IEEE, International Association of Computer Science and Information Technology (IACSIT) and International Association of Engineers (IAENG). He has given invited lectures at AICTE and TEQIP sponsored workshops and under faculty development programs, as a supporting speaker.

 <https://orcid.org/0000-0002-3947-4292>

E-mail: lokeshcse@yahoo.co.in

Department of Computer Science and Engineering
Basaveshwar Engineering College (Autonomous)
Bagalkot, Karnataka, India



Gangadharaiah S. received the B.Eng. and M.Tech. degree in Computer Science and Engineering from Visveswaraya Technological University, Belagavi, India, in 2003 and 2011, respectively. He is currently pursuing a Ph.D. degree in Computer Science and Engineering at Visveswaraya Technological University, Belagavi, India. He has been working as an Assistant Professor at the Department of Computer Science at S.R.S.M.N. Government First Grade College at Udupi, India, since 2017. He worked as a junior Design Engineer and Software Engineer for Federal Technologies and Mindtree Pvt. Ltd. from 2004–2008, respectively. He also worked as an Assistant Professor at the Department of Information Science and Engineering at Acharya Institute of Technology, Bangalore, from 2009 to 2017. His current research interests include Internet of Things, sensor networks, and cloud computing.

 <https://orcid.org/0000-0001-9233-7740>

E-mail: gangadhar.s@gmail.com

Department of Computer Science
S.R.S.M.N. Government First Grade College
Udupi, Karnataka, India

Designing a Compact Microstrip Antenna Using the Machine Learning Approach

Kanhaiya Sharma and Ganga Prasad Pandey

School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, India

<https://doi.org/10.26636/jtit.2020.143520>

Abstract—This paper presents how machine learning techniques may be applied in the process of designing a compact dual-band H-shaped rectangular microstrip antenna (RMSA) operating in 0.75–2.20 GHz and 3.0–3.44 GHz frequency ranges. In the design process, the same dimensions of upper and lower notches are incorporated, with the centered position right in the middle. Notch length and width are verified for investigating the antenna. An artificial neural network (ANN) model is developed from the simulated dataset, and is used for shape prediction. The same dataset is used to create a mathematical model as well. The predicted outcome is compared and it is determined that the model relying on ANN offers better results.

Keywords—artificial neural network, dual band, microstrip antenna, notch.

1. Introduction

New developments in the field of wireless communications call for compact, wide-band, high efficiency, multiband, and low cost antennas that are suitable for modern day applications. Microstrip antennas offer numerous advantages, such as small size, low weight and ease of fabrication due to their planar configuration. An antenna of this type is easy to mount on a rigid surface, supports dual-polarization and, unlike solutions of other types, multiband operation. The different techniques relied upon to make the antenna compact [1] without affecting its basic parameters include reactive loading [2], using a monopole antenna with defected ground structure (DGS), and modified patch [3]–[7]. Those methods are capable of reducing the antenna size by up to 70%. Meta-material [8]–[14] and engineered ground structure [15]–[18] may reduce the size of the antenna by up to 74% as well. Other miniaturization methods, like deployment of fractal structures in antenna design [19]–[21] may reduce the size of the antenna by up to 75% and engineered substrates [22]–[24] allow to reduce the size of antenna by 80% and by introducing close-ended or open-ended slots [25]–[30] of various shapes H, L, U, E, V, W, D, S double E, double U may reduce the size of the antenna by up to 86.5%. In some cases, more than one technique is used to make the antenna as compact as possible.

Machine learning (ML) is a subset of artificial intelligence (AI), effectively used in all areas of engineering, espe-

cially in communication network automation. ML is a specific type of data analytic techniques enabling machines to learn from experience (training) and to predict new data. A comparison between the traditional and the ML approach is given in Fig. 1. ML is preferred to solve complex mathematical problems with lots of variables and massive amounts of data. Today's methods relied upon while designing antennas require such a high level of expertise, as in most cases the design process requires a trade-off between such antenna parameters as small size and good bandwidth. AI approaches have been used in antenna design and optimization since the 1990s [31]. The most significant advantages that the use of the AI approach in antenna design has over the traditional methods lies in the ability to simultaneously handle multi-objective problems in order to achieve the specified goal and to provide the best automated solution [32]–[35].

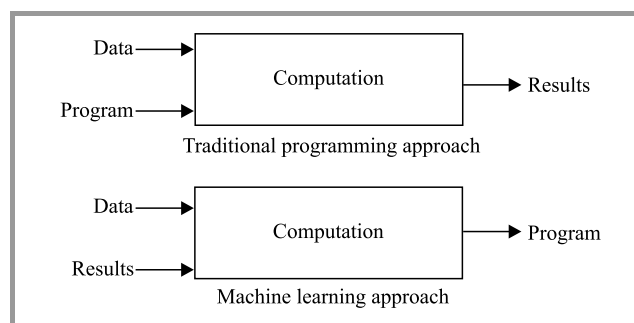


Fig. 1. Computational approaches.

The drawback of analytical-based methods consists in big-computation overheads experienced in the design process. To cope with this, the trial-and-error approach may be used to achieve the specified goals. Slot-loading turns out to be the technique that is most widely used among all aforementioned methods allowing to make antennas compact [28] in size.

2. Synthesis and Analysis Problem for Microstrip Patch Antenna Using ANN

ANN is one of the approaches used in ML to map nonlinear data efficiently [36], [37] based on experience (training),

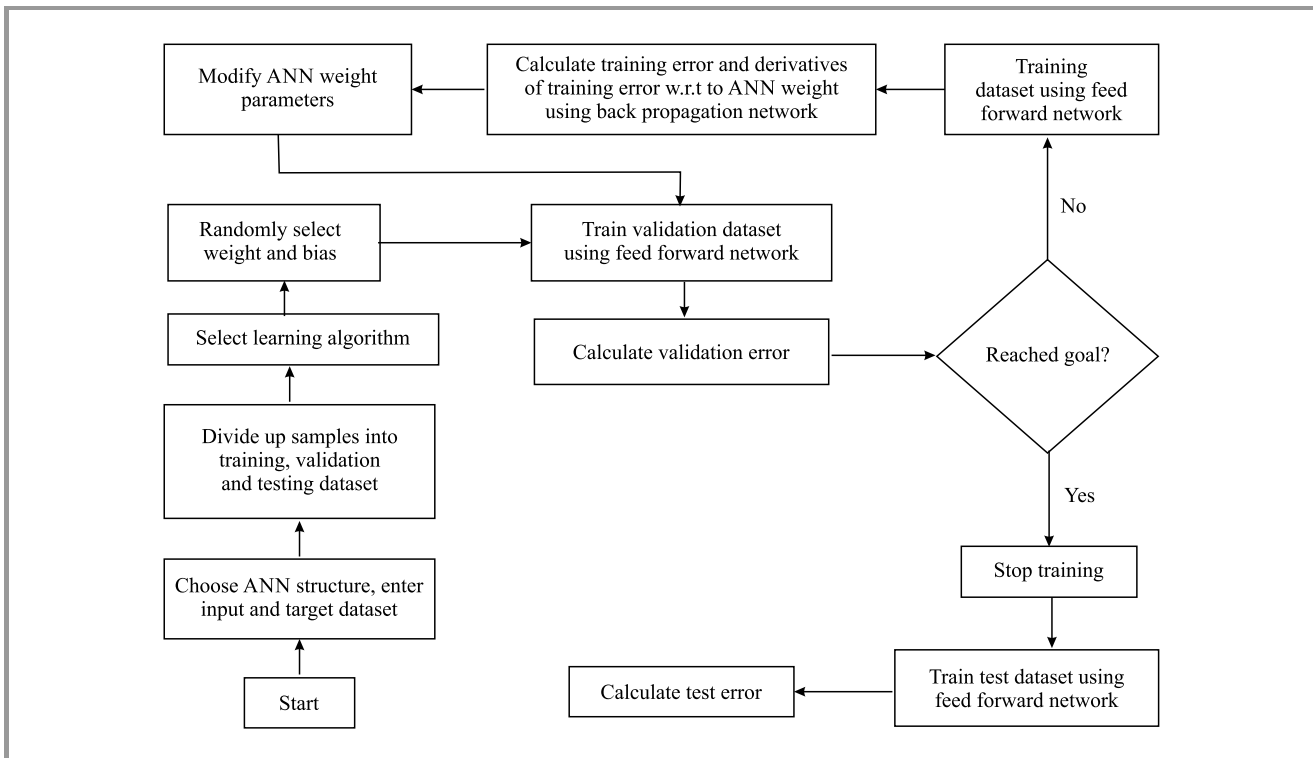


Fig. 2. Flowchart of ANN back-propagation algorithm.

while dealing with new data. Basically, three logic layers are distinguished in ANN architectures. The first is known as the input layer, while the last is referred to as the output layer. Between the input and the output layer, there is a number of hidden layers. Each hidden layer has one or more neurons. The number of the hidden layers, neurons, the activation function and the learning algorithm are all specific to the application. There are no rules to determine the number of hidden layers and neurons, but in most problems two to three hidden layers are used at the most in order to approximate all types of mathematical functions. The performance of ANN model depends upon the data collection, learning algorithm, weight initialization, change in an activation function, etc. In the case of antenna design, data should be collected either through simulations or measurements. The range of data always extends, marginally, beyond the model's utilization range [38]. Initially, data samples are divided into three sets, known as training (with the usage ratio of 70%), testing (15%), and validation (15%). According to the needs of a specific application, the percentages may be different. Next, the network size is chosen, i.e. the number of hidden layers and neurons in each of them is determined. Finally, the algorithm is selected based on feed-forward, back propagation, and feed-forward back propagation, to train the ANN and obtain the model. A flowchart of ANN feed-forward back-propagation algorithms is presented in Fig. 2.

To achieve the minimum mean squared error (MSE) It is necessary to train the model with accurate data. In this case, a total of 80 samples has been collected by simulating H-shaped RMSA using IE3D software. To minimize

MSE, different ANN algorithms have been tested by varying the number of hidden layers and neurons in each hidden layer. Finally, the best combination for the proposed model was selected. To train the ANN model, highly accu-

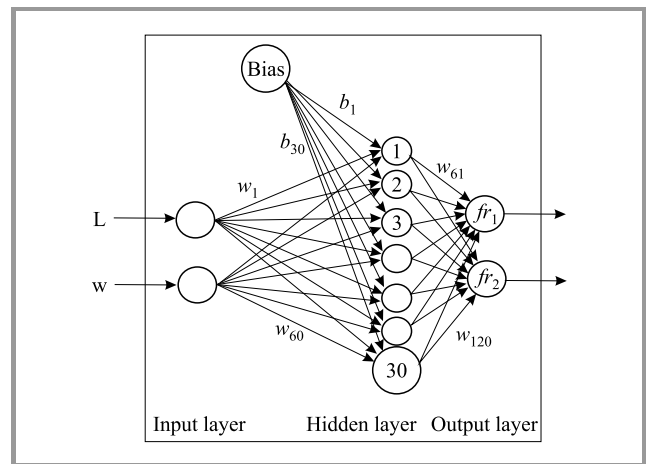


Fig. 3. Proposed ANN model for the analysis of compact H-shaped RMSA.

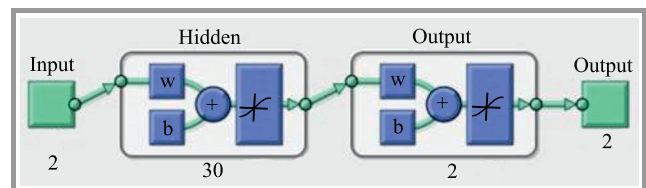


Fig. 4. ANN network architecture to train physical-to-electrical parameter.

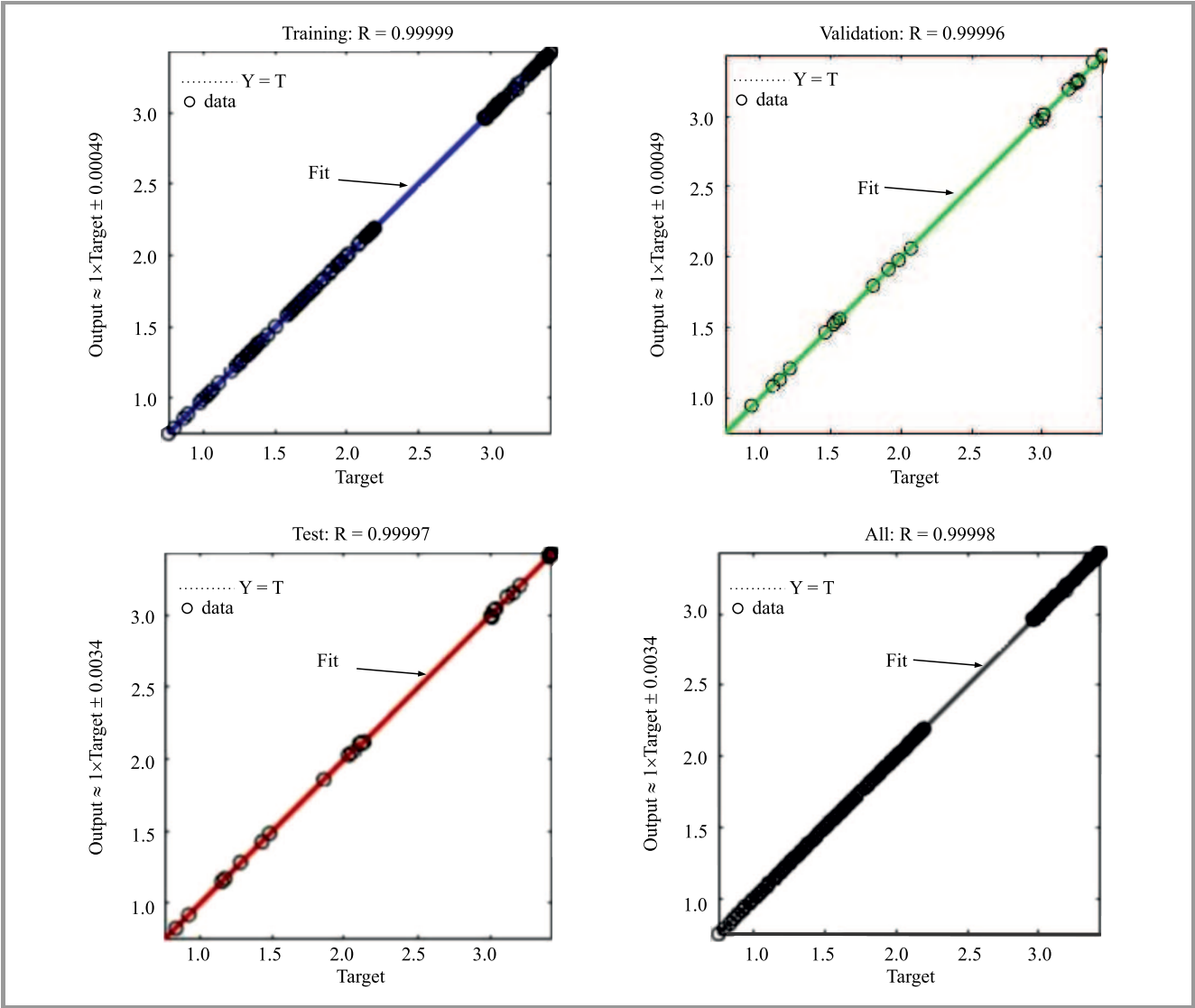


Fig. 5. Regression plot for physical to electrical parameters.

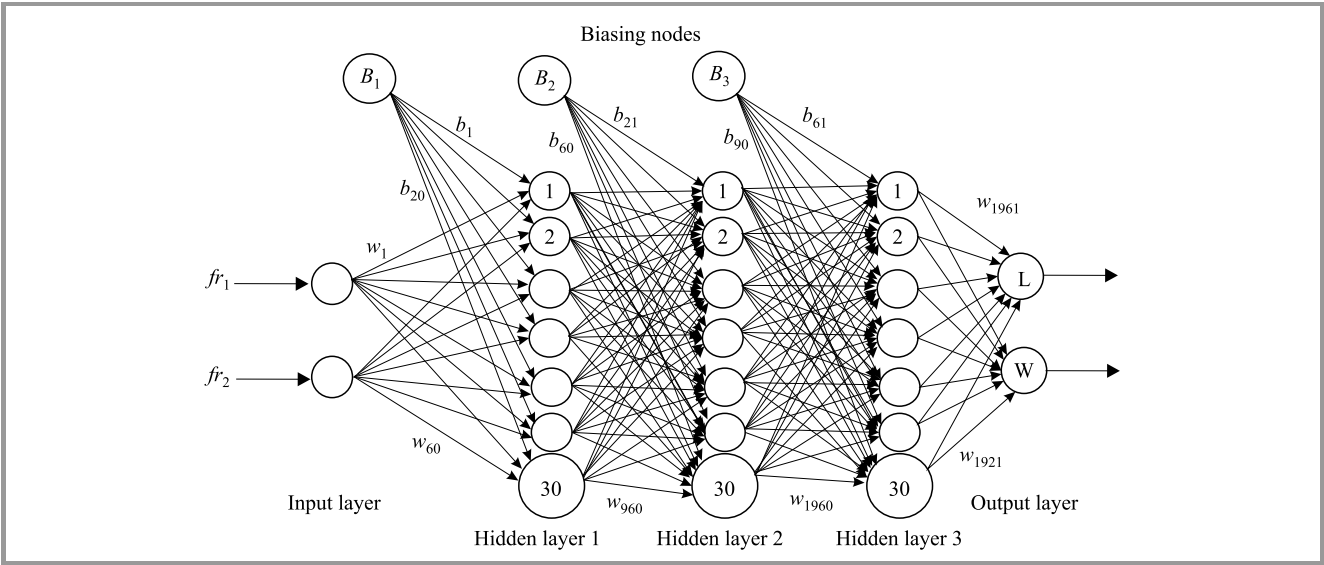


Fig. 6. Proposed ANN model for synthesis of H-shaped compact RMSA.

rate back propagation algorithms with minimum MSE were used [39].

Figure 3 shows a detailed architecture of the hidden layer of the proposed ANN model which uses the train set and predicts the dual resonant frequencies (f_{r1}, f_{r2}), as well as the physical aerial parameters, i.e. slot length (L) and width (W) as input values. The Levenberg-Marquardt (LM) back propagation algorithm is used as ANN. ANN network architecture is shown in Fig. 4 with two input and output parameters, one hidden layer and 30 neurons. The regression plot of training, testing, validation, and overall process is shown in Fig. 5.

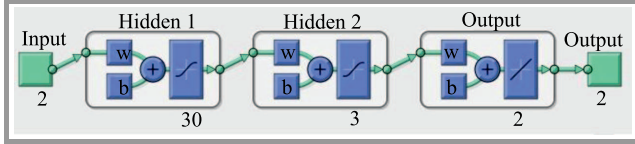


Fig. 7. ANN network architecture to train electrical-to-physical parameters.

To train the ANN model for predicting physical dimensions of antenna, Bayesian Regularization (BR) back propagation algorithm have been used. ANN network architecture is shown in Figs. 6–7, where input and output parameters equals to two, the number of hidden layer is three, each hidden layer consists of thirty neurons. the regression plot of training, testing, and overall process is shown in Fig. 8. In the next step, using the same dataset, a mathematical model was developed to map the dominant resonance mode (TM_{10}) and the higher mode (TM_{02}) of a dual band compact H-shaped patch antenna using the curve fitting tool in Matlab software. Equation 1 represents the dominant mode (TM_{10}) and Eq. 2 represents the higher mode (TM_{02}):

$$fr_1 = 2.279 - 0.145l + .07311w, \quad (1)$$

$$fr_2 = 3.37 + 0.1202l - 0.081w - 0.05906l^2 + 0.04877lw + 0.002621l^3 - 0.002256l^2w. \quad (2)$$

3. Result and Discussion

Analysis of the H-shaped compact RMSA using the proposed ANN model is performed by taking one hidden layer and 30 neurons, as shown in Fig. 4. 80 samples were taken to train the model, with 70% used for training, 15% for validation and 15% for testing. Figure 5 shows the regression plots for training, validation, testing, and the overall process. The regression R value indicating the correlation between output and target, R value closer to one means that the close relationship. In the ideal case, both outputs and target are equal. It is clear from the evidence shown that the R value for training, validation, and testing of the proposed model is 99.998% accurate, meaning that it is perfect regression model. The synthesis of H-shaped compact RMSA using the proposed ANN model is performed by taking three hidden layers and 30 neurons per hidden layer, as shown in ANN network architecture in Figs. 6–7.

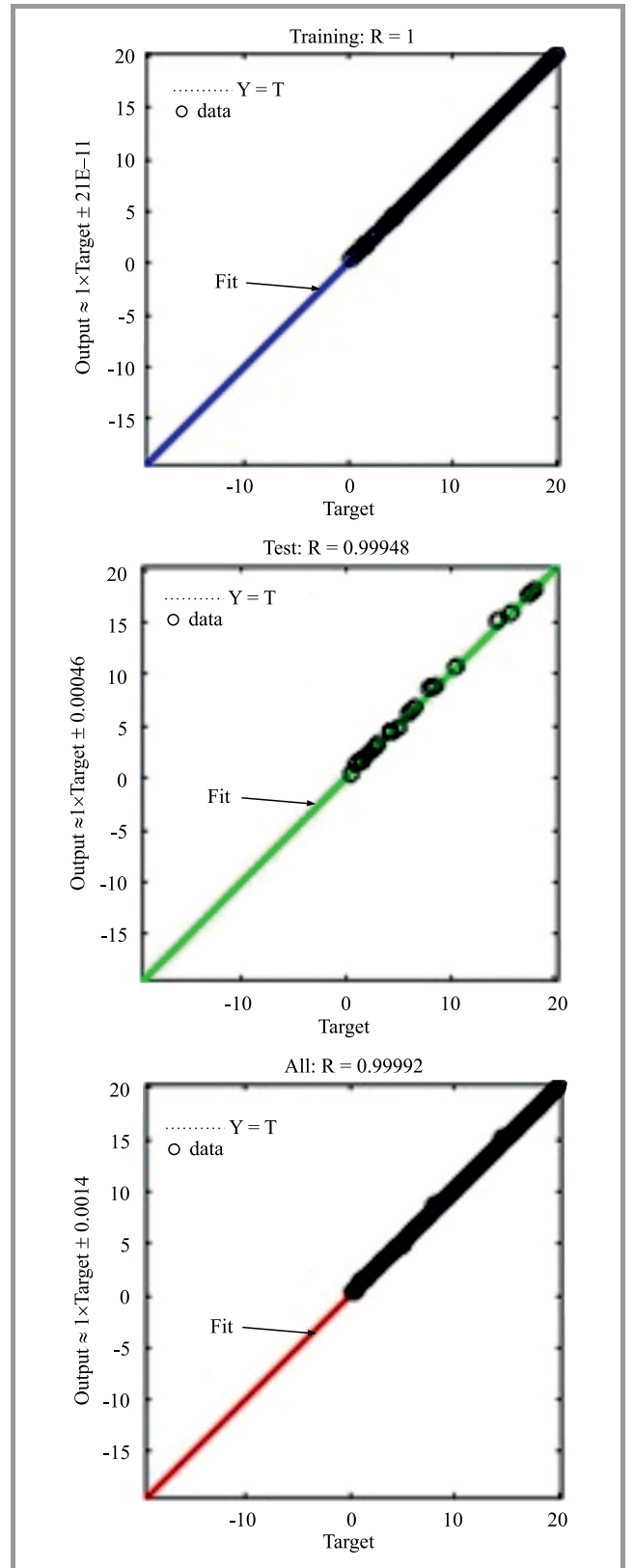


Fig. 8. Regression plot for electrical to a physical parameters.

In this case, 80 samples are used to train the model (85% for training, 15% for testing). Figure 8 shows regression plots for training, testing, and the overall process. It is clear from the evidence shown that the regression R value

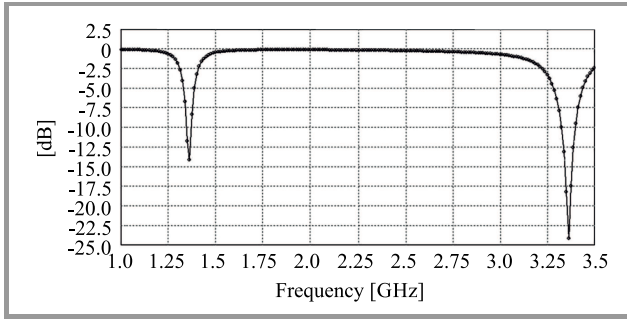


Fig. 9. S_{11} vs. frequency plot for H-shaped RMSA antenna.

for training and testing of the proposed model is 99.992%, which is very close to benchmark value 1. In the ideal scenario, both output and target are equal. In both the cases, the variations are placed precisely on dotted line, as shown in Figs. 5 and 8.

Table 1 shows the result of the proposed ANN model used for analyzing a compact H-shaped RMSA. Except for a few samples, MSE is of the order of 10^{-5} , which indicates that the predicted resonant frequencies are close to simulated results (f_{r1}, f_{r2}). Table 2 shows that the MSE is of the order of 10^{-12} , which is very close to actual physical slot dimensions. Table 3 shows a comparison of f_{r1}, f_{r2} collected from simulation, curve fitting, and ANN prediction. The dual resonant frequencies predicted by the proposed ANN model are very close to the simulated ones. The relation between rectangular notch dimensions and resonant frequencies f_{r1} and f_{r2} for the dominant mode (TM_{10} and for the higher mode TM_{02}) are shown in Fig. 10. The dimension of notches is increased from 0.25×0.5 mm to 20.00×20.25 mm, in 0.25 mm steps. f_{r1} started falling from 2.20 to 0.75 GHz, while f_{r2} first increased from 3.38 GHz to 3.434 GHz, and then suddenly decreased down to 2.98 GHz, and started increasing again from 2.98 to 3.19 GHz.

To validate the performance of H-shaped dual-band RMSA, one sample was chosen randomly with the same dimensions as those shown in Table 1, with the notch size of 8×14 mm, as well as with identical lower and upper notches centered in the middle. The infinite substrate and the infinite ground plane were assumed and simulated using IE3D software. The performance of the antenna measured in terms of scattering parameter, gain and radiation pattern was analyzed. S_{11} vs. frequency characteristics shown in Fig. 9, indicate that the antenna radiates at 1.360 GHz and 3.366 GHz for two modes (TM_{10} and TM_{02}). From Fig. 11, one may conclude that the gain of the antenna in question measured at 1.360 GHz is 7.572 dBi, and equals, at the upper resonant frequency of 3.366 GHz, 1.00 dBi. E-plane and H-plane radiation patterns at 1.360 GHz and 3.366 GHz with infinite substrate and infinite ground plane dimensions (Fig. 12) show TM_{10} and TM_{02} modes of operation, which is quite normal for any MSA. The proposed aerial offers better results in terms of accuracy and time complexity, with a higher

Table 1
Physical-to-electrical parameters

Prediction of physical parameters (slot length, slot width) to electrical (f_{r1}, f_{r2}) data						
Slot (l, w)	Simulation		ANN prediction		ANN model error	
L, W	f_{r1}	f_{r2}	f_{r1}	f_{r2}	f_{r1}	f_{r2}
0.25, 0.5	2.19209	3.38418	2.19419	3.37800	-2.10E-03	6.18E-03
0.5, 0.75	2.19128	3.38579	2.19032	3.37734	9.58E-04	8.45E-03
0.75, 1	2.18725	3.36158	2.18592	3.37702	1.33E-03	-1.54E-02
1, 1.25	2.17918	3.36965	2.18099	3.37718	-1.81E-03	-7.53E-03
1.25, 1.5	2.17918	3.36965	2.17556	3.37798	3.62E-03	-8.33E-03
1.5, 1.75	2.17111	3.38579	2.16963	3.37965	1.48E-03	6.14E-03
1.75, 2	2.159	3.38176	2.16319	3.38238	-4.19E-03	-6.21E-04
2, 2.25	2.15819	3.38983	2.15618	3.38633	2.01E-03	3.50E-03
2.25, 2.5	2.14689	3.38983	2.14848	3.39150	-1.59E-03	-1.67E-03
2.5, 2.75	2.13882	3.3979	2.13989	3.39770	-1.07E-03	2.00E-04
2.75, 3	2.13075	3.40194	2.13019	3.40449	5.58E-04	-2.55E-03
3, 3.25	2.12268	3.41001	2.11927	3.41125	3.41E-03	-1.24E-03
3.25, 3.5	2.10734	3.41808	2.10715	3.41736	1.94E-04	7.18E-04
3.5, 3.75	2.0904	3.42373	2.09398	3.42232	-3.58E-03	1.41E-03
3.75, 4	2.0791	3.42938	2.07995	3.42589	-8.54E-04	3.49E-03
4, 4.25	2.0678	3.42938	2.06515	3.42813	2.65E-03	1.25E-03
4.25, 4.5	2.05198	3.42147	2.04960	3.42924	2.38E-03	-7.77E-03
4.5, 4.75	2.02986	3.43422	2.03329	3.42955	-3.43E-03	4.67E-03
4.75, 5	2.0113	3.42933	2.01622	3.42935	-4.92E-03	-2.47E-05
5, 5.25	1.99435	3.42373	1.99848	3.42897	-4.13E-03	-5.24E-03
5.25, 5.5	1.98144	3.43019	1.98024	3.42870	1.20E-03	1.49E-03
5.5, 5.75	1.96045	3.42938	1.96183	3.42879	-1.38E-03	5.88E-04
5.75, 6	1.94915	3.43019	1.94733	3.42941	5.42E-03	7.82E-04
6, 6.25	1.92655	3.42938	1.92650	3.43041	4.79E-05	-1.03E-03
6.25, 6.5	1.9096	3.43503	1.91041	3.43109	-8.10E-04	3.94E-03
6.5, 6.75	1.89266	3.42615	1.89493	3.42987	-2.27E-03	-3.72E-03
6.75, 7	1.88055	3.42211	1.87854	3.42462	2.01E-03	-2.51E-03
7, 7.25	1.85876	3.41243	1.85964	3.41401	-8.76E-04	-1.58E-03
7.25, 7.5	1.83616	3.40678	1.83835	3.39919	-2.19E-03	7.59E-03
7.5, 7.75	1.81356	3.40113	1.81694	3.38378	-3.38E-03	1.73E-02
7.75, 8	1.79984	3.36158	1.79758	3.37135	2.26E-03	-9.77E-03
8, 8.25	1.77966	3.36965	1.78009	3.36310	-4.29E-04	6.55E-03
8.25, 8.5	1.75706	3.36723	1.76202	3.35787	-4.96E-03	9.36E-03
8.5, 8.75	1.73931	3.35755	1.74105	3.35359	-1.74E-03	3.96E-03
8.75, 9	1.71913	3.34948	1.71722	3.34863	1.91E-03	8.49E-04
9, 9.25	1.69492	3.34463	1.69299	3.34215	1.93E-03	2.48E-03
9.25, 9.5	1.67232	3.32768	1.67094	3.33390	1.38E-03	-6.22E-03
9.5, 9.75	1.65052	3.32123	1.65179	3.32382	-1.27E-03	-2.59E-03
9.75, 10	1.63035	3.30912	1.63440	3.31195	-4.05E-03	-2.83E-03
10, 10.25	1.61824	3.29701	1.61715	3.29858	1.09E-03	-1.57E-03
10.25, 10.5	1.59887	3.28249	1.59908	3.28440	-2.10E-04	-1.91E-03
10.5, 10.75	1.58192	3.26877	1.58012	3.27032	1.80E-03	-1.55E-03
10.75, 11	1.56174	3.25666	1.56061	3.25704	1.13E-03	-3.84E-04
11, 11.25	1.53672	3.25424	1.54085	3.24467	-4.13E-03	9.57E-03
11.25, 11.5	1.51977	3.23729	1.52096	3.23276	-1.19E-03	4.53E-03
11.5, 11.75	1.50121	3.22437	1.50100	3.22060	2.11E-04	3.77E-03
11.75, 12	1.48103	3.20823	1.48105	3.20739	-2.26E-05	8.39E-04
12, 12.25	1.46086	3.18805	1.46127	3.19248	-4.14E-04	-4.43E-03
12.25, 12.5	1.44633	3.16949	1.44182	3.17556	4.51E-03	-6.07E-03
12.5, 12.75	1.42857	3.15981	1.42277	3.15689	5.80E-03	2.92E-03
12.75, 13	1.40839	3.13963	1.40409	3.13736	4.30E-03	2.27E-03
13, 13.25	1.38015	3.11945	1.38564	3.11828	-5.49E-03	1.17E-03
13.25, 13.5	1.36723	3.09605	1.36729	3.10072	-5.51E-05	-4.67E-03
13.5, 13.75	1.3519	3.08717	1.34903	3.08499	2.87E-03	2.18E-03
13.75, 14	1.33172	3.07103	1.33100	3.07076	7.23E-04	2.67E-04
14, 14.25	1.31961	3.05892	1.31328	3.05773	6.33E-03	1.19E-03
14.25, 14.5	1.29944	3.0565	1.29588	3.04595	3.56E-03	1.05E-02
14.5, 14.75	1.27926	3.03874	1.27872	3.03576	5.38E-04	2.98E-03
14.75, 15	1.25989	3.02825	1.26165	3.02751	-1.76E-03	7.44E-04
15, 15.25	1.25101	3.03874	1.24446	3.02143	6.55E-03	1.73E-02
15.25, 15.5	1.22599	3.01695	1.22687	3.01752	-8.80E-04	-5.65E-04
15.5, 15.75	1.20904	3.0113	1.20869	3.01522	3.49E-04	-3.92E-03
15.75, 16	1.19209	3.0113	1.18994	3.01306	2.15E-03	-1.76E-03
16, 16.25	1.1703	3.00646	1.17083	3.00867	-5.34E-04	-2.21E-03
16.25, 16.5	1.15012	3.01049	1.15145	3.00000	-1.33E-03	1.05E-02
16.5, 16.75	1.13801	2.99839	1.13131	2.98718	6.70E-03	1.12E-02
16.75, 17	1.10169	2.95803	1.10966	2.97337	-7.97E-03	-1.53E-02
17, 17.25	1.08959	2.96207	1.08634	2.96319	3.25E-03	-1.12E-03
17.25, 17.5	1.06215	2.96045	1.06208	2.95967	6.62E-05	7.76E-04
17.5, 17.75	1.03955	2.9661	1.03784	2.96245	1.71E-03	3.65E-03
17.75, 18	1.01695	2.9774	1.01407	2.96922	2.88E-03	8.18E-03
18, 18.25	0.98870	2.97821	0.99068	2.97837	-1.98E-03	-1.62E-04
18.25, 18.5	0.97175	2.9887	0.96723	2.99017	4.52E-03	-1.47E-03
18.5, 18.75	0.93624	3.00646	0.94302	3.00600	-6.79E-03	4.57E-04
18.75, 19	0.92009	3.02663	0.91728	3.02733	2.81E-03	-7.04E-04
19, 19.25	0.88701	3.05085	0.88933	3.05488	-2.33E-03	-4.03E-03
19.25, 19.5	0.85876	3.0904	0.85881	3.08823	-5.32E-05	2.17E-03
19.5, 19.75	0.83051	3.11864	0.82578	3.12581	4.73E-03	-7.17E-03
19.75, 20	0.79096	3.18805	0.79066	3.16525	2.95E-04	2.28E-02
20, 20.25	0.75141	3.19209	0.75411	3.20409	-2.70E-03	-1.20E-02

Table 2
Electrical-to-physical parameters

Slot (l, w)		Prediction of electrical parameters (f_{r1}, f_{r2}) to physical parameters (slot length, slot width)					
L	W	Input to ANN		ANN prediction		Error in prediction	
		f_{r1}	f_{r2}	f_{r1}	f_{r2}	f_{r1}	f_{r2}
0.25	0.5	2.19209	3.38418	0.25000	0.50000	4.26E-11	4.26E-11
0.5	0.75	2.19128	3.38579	0.23763	0.48763	0.26237	0.26237
0.75	1	2.18725	3.36158	0.75000	1.00000	7.43E-11	7.43E-11
1	1.25	2.17918	3.36965	1.25000	1.50000	-0.25	-0.25
1.25	1.5	2.17918	3.36965	1.25000	1.50000	5.21E-11	5.21E-11
1.5	1.75	2.17111	3.38579	1.50000	1.75000	5.05E-11	5.05E-11
1.75	2	2.15900	3.38176	1.75000	2.00000	5.57E-11	5.57E-11
2	2.25	2.15819	3.38983	2.05693	2.30693	-0.05693	-0.05693
2.25	2.5	2.14689	3.38983	2.25000	2.50000	5.40E-11	5.40E-11
2.5	2.75	2.13882	3.39790	2.50000	2.75000	4.79E-11	4.79E-11
2.75	3	2.13075	3.40194	2.71920	2.96920	0.03080	0.03080
3	3.25	2.12268	3.41001	3.11056	3.36056	-0.11056	-0.11056
3.25	3.5	2.10734	3.41808	3.25000	3.50000	4.22E-11	4.22E-11
3.5	3.75	2.09040	3.42373	3.50000	3.75000	3.58E-11	3.58E-11
3.75	4	2.07910	3.42938	3.75000	4.00000	3.03E-11	3.03E-11
4	4.25	2.06780	3.42938	4.00000	4.25000	2.51E-11	2.51E-11
4.25	4.5	2.05198	3.42147	4.25000	4.50000	2.95E-11	2.95E-11
4.5	4.75	2.02986	3.43422	4.50000	4.75000	1.48E-11	1.48E-11
4.75	5	2.01130	3.42933	4.75000	5.00000	1.46E-11	1.46E-11
5	5.25	1.99435	3.42373	4.64830	4.89830	0.35170	0.35170
5.25	5.5	1.98144	3.43019	5.25000	5.50000	8.55E-12	8.55E-12
5.5	5.75	1.96045	3.42938	5.50000	5.75000	2.45E-12	2.45E-12
5.75	6	1.94915	3.43019	5.75000	6.00000	4.59E-12	4.59E-12
6	6.25	1.92655	3.42938	6.07401	6.32401	-0.07400	-0.07400
6.25	6.5	1.90960	3.43503	6.25000	6.50000	1.19E-11	1.19E-11
6.5	6.75	1.89266	3.42615	6.55910	6.80910	-0.05909	-0.05909
6.75	7	1.88055	3.42211	6.75000	7.00000	2.72E-11	2.72E-11
7	7.25	1.85876	3.41243	7.00000	7.25000	3.61E-11	3.61E-11
7.25	7.5	1.83616	3.40678	7.25000	7.50000	4.98E-11	4.98E-11
7.5	7.75	1.81356	3.40113	7.50000	7.75000	5.88E-11	5.88E-11
7.75	8	1.79984	3.36158	7.75000	8.00000	-6.34E-12	-6.35E-12
8	8.25	1.77966	3.36965	8.51865	8.76865	-0.51865	-0.51865
8.25	8.5	1.75706	3.36723	8.25000	8.50000	-3.08E-12	-3.08E-12
8.5	8.75	1.73931	3.35755	8.62045	8.87045	-0.12045	-0.12045
8.75	9	1.71913	3.34948	8.75000	9.00000	-7.34E-12	-7.34E-12
9	9.25	1.69492	3.34463	9.00000	9.25000	-9.65E-12	-9.65E-12
9.25	9.5	1.67232	3.32768	9.25000	9.50000	-9.32E-12	-9.32E-12
9.5	9.75	1.65052	3.32123	9.50000	9.75000	-1.20E-11	-1.20E-11
9.75	10	1.63035	3.30912	9.75000	10.0000	-9.44E-12	-9.44E-12
10	10.25	1.61824	3.29701	10.0000	10.2500	-1.13E-11	-1.13E-11
10.25	10.5	1.59887	3.28249	10.2500	10.5000	-2.71E-12	-2.71E-12
10.5	10.75	1.58192	3.26877	10.5246	10.7746	-0.02457	-0.02457
10.75	11	1.56174	3.25666	10.7500	11.0000	-6.22E-12	-6.22E-12
11	11.25	1.53672	3.25424	11.0000	11.2500	-7.51E-12	-7.51E-12
11.25	11.5	1.51977	3.23729	11.2500	11.5000	1.12E-11	1.12E-11
11.5	11.75	1.50121	3.22437	11.5000	11.7500	-1.75E-11	-1.75E-11
11.75	12	1.48103	3.20823	11.7500	12.0000	8.74E-12	8.74E-12
12	12.25	1.46086	3.18805	12.0000	12.2500	-3.49E-12	-3.49E-12
12.25	12.5	1.44633	3.16949	12.2500	12.5000	3.14E-12	3.14E-12
12.5	12.75	1.42857	3.15981	12.5000	12.7500	-1.12E-12	-1.12E-12
12.75	13	1.40839	3.13963	12.7500	13.0000	5.51E-13	5.51E-13
13	13.25	1.38015	3.11945	13.0000	13.2500	7.48E-13	7.48E-13
13.25	13.5	1.36723	3.09605	13.2500	13.5000	-4.42E-13	-4.42E-13
13.5	13.75	1.35190	3.08717	13.5000	13.7500	1.74E-13	1.74E-13
13.75	14	1.33172	3.07103	13.7500	14.0000	-2.08E-12	-2.08E-12
14	14.25	1.31961	3.05892	14.0000	14.2500	-4.21E-12	-4.21E-12
14.25	14.5	1.29944	3.0565	14.2500	14.5000	-6.14E-12	-6.14E-12
14.5	14.75	1.27926	3.03874	14.9728	15.2228	-0.47284	-0.47284
14.75	15	1.25989	3.02825	14.7500	15.0000	-1.61E-10	-1.61E-10
15	15.25	1.25101	3.03874	15.0000	15.2500	-1.07E-10	-1.07E-10
15.25	15.5	1.22599	3.01695	15.2500	15.5000	-1.07E-10	-1.07E-10
15.5	15.75	1.20904	3.0113	15.5000	15.7500	-6.79E-11	-6.79E-11
15.75	16	1.19209	3.01130	15.6709	15.9209	0.07901	0.07901
16	16.25	1.17030	3.00646	16.0000	16.2500	-4.76E-11	-4.76E-11
16.25	16.5	1.15012	3.01049	16.2500	16.5000	-7.06E-11	-7.06E-11
16.5	16.75	1.13801	2.99839	16.5000	16.7500	-2.86E-11	-2.86E-11
16.75	17	1.10169	2.95803	16.7500	17.0000	-5.46E-12	-5.46E-12
17	17.25	1.08959	2.96207	17.0000	17.2500	-5.68E-12	-5.68E-12
17.25	17.5	1.06215	2.96045	17.2500	17.5000	-4.63E-12	-4.63E-12
17.5	17.75	1.03955	2.96610	17.4709	17.7209	0.02905	0.02906
17.75	18	1.01695	2.97740	17.7500	18.0000	-8.48E-12	-8.48E-12
18	18.25	0.98870	2.97821	17.8963	18.1463	0.10368	0.10368
18.25	18.5	0.97175	2.98870	18.2500	18.5000	-2.24E-11	-2.24E-11
18.5	18.75	0.93624	3.00646	18.5000	18.7500	-1.19E-10	-1.19E-10
18.75	19	0.92010	3.02663	18.7500	19.0000	-3.99E-10	-3.99E-10
19	19.25	0.88701	3.05085	19.0000	19.2500	-4.97E-11	-4.97E-11
19.25	19.5	0.85876	3.09040	19.2500	19.5000	-2.32E-11	-2.32E-11
19.5	19.75	0.83051	3.11864	19.5000	19.7500	-1.22E-11	-1.22E-11
19.75	20	0.79096	3.18805	19.7500	20.0000	1.98E-11	1.98E-11
20	20.25	0.75141	3.19209	20.0000	20.2500	1.50E-11	1.50E-11

Table 3
Comparison of resonant frequencies

Composion of resonant frequencies in various methods							
Slot size		Simulation		Curve fitting		ANN model	
L	W	f ₁	f ₂	f ₁	f ₂	f ₁	f ₂
0.25	0.5	2.19209	3.38418	2.2793	3.3619	2.19323512	3.383955179
0.5	0.75	2.19128	3.38579	2.2613	3.3728	2.190414971	3.377723607
0.75	1	2.18725	3.36158	2.2434	3.3823	2.186759462	3.373556577
1	1.25	2.17918	3.36965	2.2254	3.3907	2.18221719	3.37167556
1.25	1.5	2.17918	3.36965	2.2074	3.3977	2.176752125	3.372181943
1.5	1.75	2.17111	3.38579	2.1894	3.4036	2.170350743	3.374991672
1.75	2	2.159	3.38176	2.1751	3.4084	2.163026757	3.379792164
2	2.25	2.15819	3.38983	2.1535	3.412	2.154820328	3.386051178
2.25	2.5	2.14689	3.38983	2.1355	3.4146	2.145790119	3.393092756
2.5	2.75	2.13882	3.3979	2.1176	3.4161	2.136000000	3.400223563
2.75	3	2.13075	3.40194	2.0996	3.4166	2.12550524	3.406863738
3	3.25	2.12268	3.41001	2.0816	3.4161	2.11434336	3.412633595
3.25	3.5	2.10734	3.41808	2.0636	3.4147	2.102532065	3.417373395
3.5	3.75	2.0904	3.42373	2.0457	3.4123	2.090073149	3.421106618
3.75	4	2.0791	3.42938	2.0277	3.4091	2.076959378	3.423975275
4	4.25	2.0678	3.42938	2.0097	3.405	2.063181506	3.426174076
4.25	4.5	2.05198	3.42147	1.9917	3.4001	2.048733878	3.427898365
4.5	4.75	2.02986	3.43422	1.9738	3.3945	2.033618293	3.429309015
4.75	5	2.0113	3.42933	1.9558	3.3881	2.017846529	3.430510785
5	5.25	1.99435	3.42373	1.9378	3.381	2.001442149	3.431538579
5.25	5.5	1.98144	3.43019	1.9199	3.3732	1.984442103	3.432346835
5.5	5.75	1.96045	3.42938	1.9019	3.3648	1.966898343	3.432799628
5.75	6	1.94915	3.43019	1.8839	3.3558	1.948879197	3.432662761
6	6.25	1.92655	3.42938	1.8659	3.3462	1.930469403	3.431604735
6.25	6.5	1.9096	3.43503	1.848	3.3361	1.911766619	3.429221116
6.5	6.75	1.89266	3.42615	1.83	3.3255	1.892871285	3.425103028
6.75	7	1.88055	3.42211	1.812	3.3144	1.873867609	3.418964553
7	7.25	1.85876	3.41243	1.794	3.3028	1.854798692	3.410809045
7.25	7.5	1.83616	3.40678	1.7761	3.2909	1.835648248	3.401051244
7.5	7.75	1.81356	3.40113	1.7581	3.2786	1.81634694	3.390475111
7.75	8	1.79984	3.36158	1.7401	3.266	1.796810053	3.379982051
8	8.25	1.77966	3.36965	1.7222	3.2531	1.776988061	3.370251304
8.25	8.5	1.75706	3.36723	1.7042	3.2399	1.756898836	3.361519085
8.5	8.75	1.73931	3.35755	1.6862	3.2265	1.736662571	3.353579746
8.75	9	1.71913	3.34948	1.6682	3.2129	1.716290452	3.345948928
9	9.25	1.69492	3.34463	1.6503	3.1992	1.69601866	3.33806789
9.25	9.5	1.67232	3.32768	1.6323	3.1853	1.675911158	3.329468411
9.5	9.75	1.65052	3.32123	1.6143	3.1713	1.656025072	3.319872427
9.75	10	1.63035	3.30912	1.5964	3.1573	1.636366626	3.309226066
10	10.25	1.61824	3.29701	1.5784	3.1433	1.616895951	3.3007672933
10.25	10.5	1.59887	3.28249	1.5604	3.1292	1.597542878	3.28547673
10.5	10.75	1.58192	3.26877	1.5424	3.1153	1.572829533	3.272915207
10.75	11	1.56174	3.25666	1.5245	3.1013	1.558892864	3.260178391
11	11.25	1.53672	3.25424	1.5065	3.0875	1.539500355	3.247030294
11.25	11.5	1.51977	3.23729	1.4885	3.0739	1.52005534	3.234163114
11.5	11.75	1.50121	3.22437	1.4705	3.0604	1.500592591	3.22050971
11.75	12	1.48103	3.20823	1.4526	3.0472	1.481167693	3.206053827
12	12.25	1.46086	3.18805	1.4346	3.0342	1.461844345	3.190561503
12.25	12.5	1.44633	3.16949	1.4166	3.0215	1.442682842	3.179445000
12.5	12.75	1.42857	3.15981	1.3987	3.0091	1.423731612	3.15632401
12.75	13	1.40839	3.13963	1.3807	2.9971	1.405022572	3.13805028
13	13.25	1.38015	3.11945	1.3627	2.9854	1.386570215	3.116969014
13.25	13.5	1.36723	3.09605	1.3447	2.9742	1.368373589	3.101893929
13.5	13.75	1.3519	3.08717	1.3268	2.9634	1.350419758	3.085223618
13.75	14	1.33172	3.07103	1.3088	2.9532	1.332686968	3.070378841
14	14.25	1.31961	3.05892	1.2908	2.9434	1.315145999	3.056765597
14.25	14.5	1.29944	3.0565	1.2728	2.9342	1.297758688	3.047257546
14.5	14.75	1.27926	3.03874	1.2549	2.9256	1.280473345	3.039037159
14.75	15	1.25989	3.02825	1.2369	2.9177	1.263217358	3.02731293
15	15.25	1.25101	3.03874	1.2189	2.9104	1.245887804	3.02779209
15.25	15.5	1.22599	3.01695	1.201	2.9038	1.228342041	3.023611391
15.5	15.75	1.20904	3.0113	1.183	2.8979	1.210392669	3.019365567
15.75	16	1.19209	3.0113	1.165	2.8928	1.191815554	3.014176798
16	16.25	1.1703	3.00646	1.147	2.8884	1.172383691	3.007259144
16.25	16.5	1.15012	3.01049	1.1291	2.885	1.151935686	2.9982168
16.5	16.75	1.13801	2.99839	1.1111	2.8824	1.130462957	2.98741289
16.75	17	1.10169	2.95803	1.0931	2.8806	1.108157717	2.976170793
17	17.25	1.08959	2.96207	1.0751	2.8799	1.085352868	2.966523582
17.25	17.5	1.06215	2.96045	1.0572	2.8801	1.062359012	2.960532524
17.5	17.75	1.03955	2.9661	1.0392	2.8813	1.039305812	2.959616637
17.75	18	1.01695	2.9774	1.0212	2.8835	1.016091761	2.964321303
18	18.25	0.988701	2.97821	1.0033	2.8868	0.992442652	2.974524237
18.25	18.5	0.971751	2.9887	0.9853	2.8912	0.968009009	2.989787327
18.5	18.75	0.936239	3.00646	0.9673	2.8963	0.942446684	3.009621486
18.75	19	0.920097	3.02663	0.9493	2.9035	0.915466504	3.033605786
19	19.25	0.887006	3.05085	0.9314	2.9114	0.886861758	3.061400275
19.25	19.5	0.858757	3.0904	0.9134	2.9206	0.856524855	3.092706667
19.5	19.75	0.830508	3.11864	0.8954	2.9311	0.824459185	3.127214353
19.75	20	0.79096	3.18805	0.8774	2.9429	0.790787012	3.164552035
20	20.25	0.751412	3.19209	0.8595	2.956	0.755751204	3.204255063

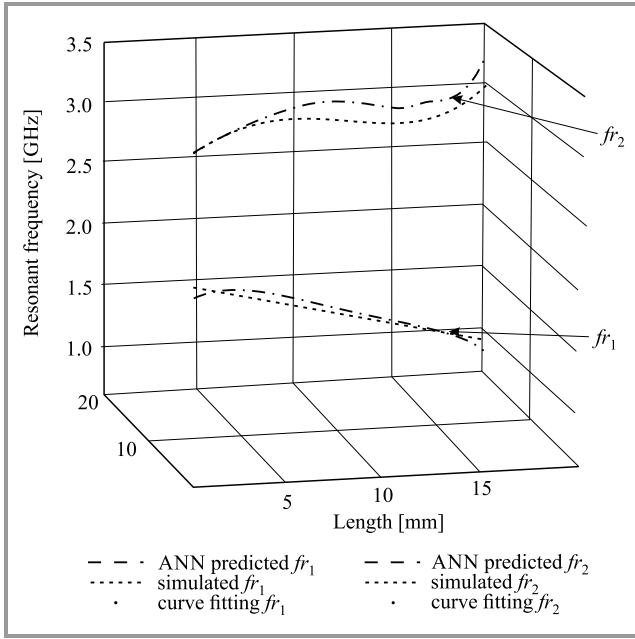


Fig. 10. Notch dimension vs. resonant frequency for TM_{10} , and TM_{02} modes of operation.

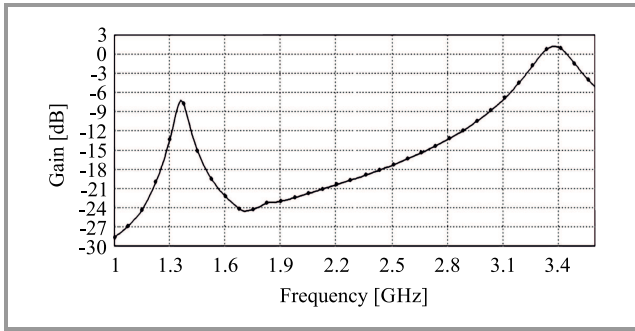


Fig. 11. Maximum gain vs. frequency plot for sample H-shaped RMSA.

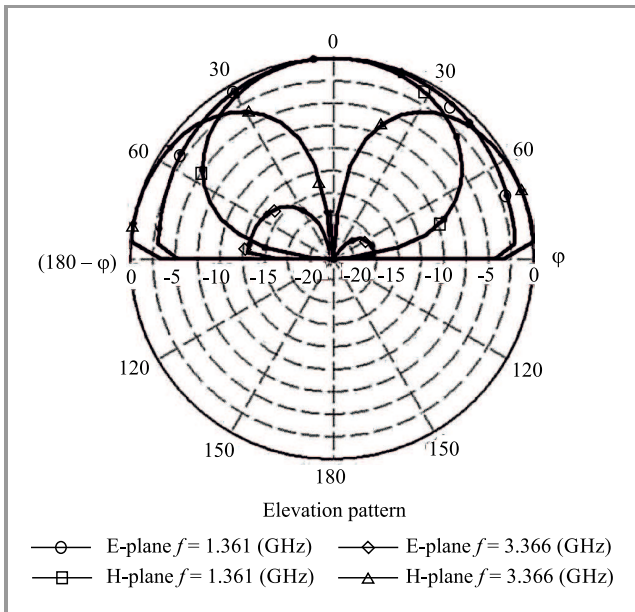


Fig. 12. Radiation pattern for sample H-shaped RMSA.

number of samples, in comparison to examples existing literature [40].

4. Design Specification and Dataset

Design specifications of the proposed dual-band H-shaped compact RMSA with an infinite ground plane are shown in Fig. 13, and the list of parameters used in the design

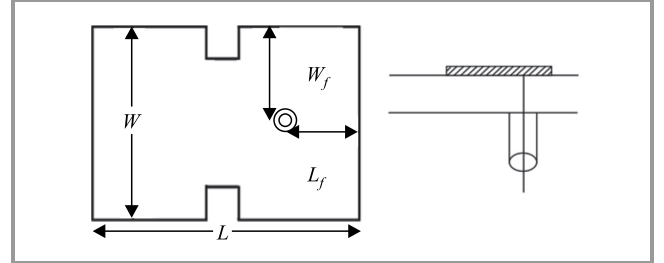


Fig. 13. Top and side view of H-shaped RMSA used for modeling.

is presented in Table 4. Along the radiating edge, two rectangular notches that are equal in dimensions are cut (upper, lower and centered in the middle), starting with the notch dimensions of 0.2×0.5 mm, increased to 20×20.25 mm in fixed steps of 0.25 mm in L and W . 80 samples were collected using IE3D simulation software.

Table 4
Specification of H-shaped RMSA

Parameter	Value
Operating frequency	2.33 GHz
Patch length	$L = 32$ mm
Patch width	$W = 42$ mm
Substrate	FR-4
Substrate thickness	1.6 mm
Dielectric constant	4.4
Loss tangent	0.025
Feed type	Coaxial
Notch length	0.25 to 20 mm
Notch width	0.5 to 20.25 mm
Feed location (L_f, W_f)	6.575 mm, 21 mm

5. Conclusion

Machine learning is one of the methods used to design antennas quickly. The proposed model is simple, time-efficient and does not require any complex mathematical calculations. It may easily predict, with high accuracy, the physical or electrical parameters within the range of data provided for training, for a dual-band (0.75 to 2.20 GHz and 3.0 to 3.44 GHz), compact microstrip antenna that may be used in a variety of wireless applications.

References

- [1] M. Fallahpour and R. Zoughi, "Antenna miniaturization techniques: A review of topology- and material-based methods", *IEEE Antenn. and Propag. Mag.*, vol. 60, no. 1, pp. 38–50, 2018 (DOI: 10.1109/MAP.2017.2774138).
- [2] P. L. Chi, R. Waterhouse, and T. Itoh, "Antenna miniaturization using slow wave enhancement factor from loaded transmission line models", *IEEE Trans. on Antenn. and Propag.*, vol. 59, no. 1, pp. 48–57, 2011 (DOI: 10.1109/TAP.2010.2090452).
- [3] A. Foudazi, H. R. Hassani, and S. M. Ali Nezhad, "Small UWB planar monopole antenna with added GPS/GSM/WLAN bands", *IEEE Trans. on Antenn. and Propag.*, vol. 60, no. 6, pp. 2987–2992, 2012 (DOI: 10.1109/TAP.2012.2194632).
- [4] V. K. Sambhe, R. N. Awale, and A. Wagh, "Compact U-shape radiating patch with rectangular ground planar monopole antenna", *The J. of Engin.*, vol. 2015, no. 2, pp. 54–58, 2015 (DOI: 10.1049/joe.2014.0269).
- [5] X. Ren, S. Gao, and Y. Yin, "Compact tri-band monopole antenna with hybrid strips for WLAN/WiMAX applications", *Microw. and Opt. Technol. Lett.*, vol. 57, no. 1, pp. 94–99, 2015 (DOI: 10.1002/mop.28785).
- [6] M.-T. Tan, B.-Z. Wang, and Z.-M. Zhang, "Compact tri-band dual-polarized planar monopole antenna with asymmetrical ground plane and loaded stub", *Progr. in Electromag. Res. Lett.*, vol. 58, pp. 89–95 2016 (DOI: 10.2528/PIERL15122004).
- [7] Y. Mao, S. Guo, and M. Chen, "Compact dual-band monopole antenna with defected ground plane for Internet of Things", *IET Microw., Antenn. Propag.*, vol. 12, no. 8, pp. 1332–1338, 2018 (DOI: 10.1049/iet-map.2017.0860).
- [8] R. O. Ouedraogo, E. J. Rothwell, A. R. Diaz, K. Fuchi, and A. Temme, "Miniaturization of patch antennas using a metamaterial-inspired technique", *IEEE Trans. on Antenn. and Propag.*, vol. 60, no. 5, pp. 2175–2182, 2012 (DOI: 10.1109/TAP.2012.2189699).
- [9] Y. Dong and T. Itoh, "Metamaterial-based antennas", *Proceedings of the IEEE*, vol. 100, no. 7, pp. 2271–2285, 2012 (DOI: 10.1109/JPROC.2012.2187631).
- [10] A. Bazrkar, A. Gudarzi, and M. Mahzoon, "Miniaturization of rectangular patch antennas partially loaded with μ -negative metamaterials", in *Proc. Int. Conf. on Electron., Biomed. Engin. and its Appl. ICEBEA'2012*, Dubai, United Arab Emirates, 2012, pp. 289–292 [Online]. Available: <http://psrcentre.org/images/extraimages/58.%200112256.pdf>
- [11] F. Raval, Y. P. Kosta, J. Makwana, and A. V. Patel, "Design implementation of reduced size microstrip patch antenna with metamaterial defected ground plane", in *Proc. of Int. Conf. on Commun. and Sig. Process.*, Melmaruvathur, India, 2013, pp. 186–190 (DOI: 10.1109/iccsp.2013.6577040).
- [12] R. A. H. Mahdi and S. M. R. Taha, "Miniaturization of rectangular microstrip patch antenna using topology optimized metamaterial", *IEICE Electron. Express*, vol. 14, no. 19, 2017 (DOI: 10.1587/ele.14.20170787).
- [13] V. G. Ajay and T. Mathew, "Size reduction of microstrip patch antenna through metamaterial approach for WiMAX application", in *Proc. of Int. Conf. on Wirel. Commun., Sig. Process. and Network. WiSPNET 2217*, Chennai, India, 2017, pp. 379–381 (DOI: 10.1109/WiSPNET.2017.8299782).
- [14] J. Chatterjee, A. Mohan, and V. Dixit, "Broadband circularly polarized H-shaped patch antenna using reactive impedance surface", *IEEE Antenn. and Wirel. Propag. Lett.*, vol. 17, no. 4, pp. 625–628, 2018 (DOI: 10.1109/LAWP.2018.2806993).
- [15] B. Gupta, S. Nakhate, and M. Shandilya, "A compact wideband microstrip patch antenna with defected ground plane", in *4th World Congr. on Inform. and Commun. Technol. WICT 2014*, Bandar Hilir, Malaysia, 2014, pp. 51–56 (DOI: 10.1109/WICT.2014.7077301).
- [16] M. K. Khandelwal, B. K. Kanaujia, S. Dwari, S. Kumar, and A. Gautam, "Analysis and design of dual band compact stacked microstrip patch antenna with defected ground structure for WLAN/WiMAX applications", *AEU – Int. J. of Electron. and Commun.*, vol. 69, no. 1, pp. 39–47, 2015 (DOI: 10.1016/j.aeue.2014.07.018).
- [17] S. Veisee, S. Asadi, and M. Hedayati, "A novel compact defected ground structure and its application in mutual coupling reduction of a microstrip antenna", *Turkish J. of Elec. Engin. and Comp. Sci.*, vol. 24, no. 5, pp. 3664–3670, 2016 (DOI: 10.3906/elk-1404-517).
- [18] P. Anitha, A. S. R. Reddy, and M. N. G. Prasad, "Design of a compact ψ -shaped micro-strip patch antenna on modified ground planes", in *Proc. of Int. Conf. on Recent Adv. in Electron. and Commun. Technol. ICRAECT 2017*, Bangalore, India, 2017, pp. 96–99 (DOI: 10.1109/ICRAECT.2017.60).
- [19] C. Sharma and V. D. Kumar, "Koch fractalized compact spiral antenna based on Fibonacci sequence", in *Proc. of IEEE 5th Asia-Pacific Conf. on Antenn. and Propag. APCAP 2016*, Kaohsiung, Taiwan, 2016, pp. 157–158 (DOI: 10.1109/APCAP.2016.7843146).
- [20] F. Wang, F. Bin, Q. Sun, J. Fan, and H. Ye, "A compact UHF antenna based on complementary fractal technique", *IEEE Access*, vol. 5, pp. 21118–21125, 2017 (DOI: 10.1109/ACCESS.2017.2756672).
- [21] Y. K. Choukiker and J. C. Mudiganti, "Compact hybrid fractal antenna for wideband wireless applications", *Int. J. of Microw. and Wirel. Technol.*, vol. 9, no. 5, pp. 1191–1196, 2017 (DOI: 10.1017/S1759078716001318).
- [22] C. Deng, Y. Li, Z. Feng, W. Li, and D. Zhang, "A dual-band circularly-polarized stacked patch antenna with a small frequency-ratio", in *Proc. IEEE Antenn. and Propag. Soc. Int. Symp. (APSURSI 2013)*, Orlando, FL, USA, 2013, pp. 942–943 (DOI: 10.1109/APS.2013.6711130).
- [23] A. Foroozesh, D. Psychoudakis, J. Homer, and I. Kim, "Compact wideband stacked microstrip patch antenna for a medical application", in *2016 IEEE Int. Symp. on Antenn. and Propag. APSURSI 2016*, Fajardo, Puerto Rico, 2016, pp. 283–284 (DOI: 10.1109/APS.2016.7695850).
- [24] A. Katyal and A. Basu, "Compact and broadband stacked microstrip patch antenna for target scanning applications", *IEEE Antenn. and Wirel. Propag. Lett.*, vol. 16, pp. 381–384, 2017 (DOI: 10.1109/LAWP.2016.2578723).
- [25] Y. Gou, S. Yang, Q. Zhu, and Z. Nie, "A compact dual-polarized double E-shaped patch antenna with high isolation", *IEEE Trans. on Antenn. and Propag.*, vol. 61, no. 8, pp. 4349–4353, 2013 (DOI: 10.1109/TAP.2013.2262664).
- [26] C. K. Hsu and S. J. Chung, "Compact antenna with U-shaped open-end slot structure for multi-band handset applications", *IEEE Trans. on Antenn. and Propag.*, vol. 62, no. 2, pp. 929–932, 2014 (DOI: 10.1109/TAP.2013.2289996).
- [27] M. Tarbouch, A. E. Amri, and H. Terchoune, "Compact CPW-fed microstrip octagonal patch antenna with h slot for WLAN and WiMAX applications", in *Proc. of Int. Conf. on Wirel. Technol., Embedd. and Intell. Syst. WITS 2017*, Fez, Morocco, 2017 (DOI: 10.1109/WITS.2017.7934638).
- [28] A. Roy, S. Bhunia, D. C. Sarkar, and P. P. Sarkar, "Slot loaded compact microstrip patch antenna for dual band operation", *Progress in Electromag. Res. C*, vol. 73, pp. 145–156, 2017 (DOI: 10.2528/PIERC17020903).
- [29] M. T. Islam, M. Samsuzzaman, and M. Z. Mahmud, "A compact slotted patch antenna for breast tumor detection", *Microw. and Opt. Technol. Lett.*, vol. 60, no. 7, pp. 1600–1608, 2018 (DOI: 10.1002/mop.31215).
- [30] Y. Li, Q. Liu, Y. Chen, C. Li, Z. Mo, and F. Li, "A compact triple wideband-notched UWB antenna", in *Proc. Int. Worksh. on Antenna Technol. iWAT 2018*, Nanjing, China, 2018 (DOI: 10.1109/IWAT.2018.8379142).
- [31] B. Choudhury, S. Thomas, and R. M. Jha, "Implementation of soft computing optimization techniques in antenna engineering [Antenna applications corner]", *IEEE Antenn. and Propag. Mag.*, vol. 57, no. 6, pp. 122–131, 2015 (DOI: 10.1109/MAP.2015.2439612).
- [32] Z. Cai, S. Zeng, Y. Y. Li, and L. Kang, "Automated antenna design using normalized steady state genetic algorithm", in *Proc. of NASA/ESA Conf. on Adapt. Hardware and Syst.*, Noordwijk, Netherlands, 2008 (DOI: 10.1109/AHS.2008.26).

- [33] S. Ledesma, J. Ruiz-Pinales, G. Cerda-Villafañá, and M. G. Garcia-Hernandez, "A hybrid method to design wire antennas: Design and optimization of antennas using artificial intelligence", *IEEE Antenn. and Propag. Mag.*, vol. 57, no. 4, pp. 23–31, 2015 (DOI: 10.1109/MAP.2015.2453912).
- [34] L. Feruglio, "Artificial intelligence for small satellites mission autonomy", Doctoral Dissertation, Politecnico di Torino, 2017 (DOI: 10.6092/polito/porto/2694565).
- [35] L. Y. Xiao, W. Shao, F. L. Jin, and B. Z. Wang, "Multiparameter modeling with ANN for antenna design", *IEEE Trans. on Antenn. and Propag.*, vol. 66, no. 7, pp. 3718–3723, 2018 (DOI: 10.1109/TAP.2018.2823775).
- [36] A. Landi, P. Piaggi, M. Laurino, and D. Menicucci, "Artificial neural networks for nonlinear regression and classification", in *Proc. of 10th In. Conf. on Intell. Syst. Design and Appl.*, Cairo, Egypt, 2010, pp. 115–120 (DOI: 10.1109/ISDA.2010.5687280).
- [37] S. Lek *et al.*, "Application of neural networks to modelling nonlinear relationships in ecology", *Ecological Modell.*, vol. 90, no. 1, 1996, pp. 39–52 (DOI: 10.1016/0304-3800(95)00142-5).
- [38] C. G. Christodoulou and A. Patnaik, "Neural networks for antennas", in *Morden Antenna Handbook*, C. A. Balanis, Ed. Wiley, 2008, pp. 1625–1657 (ISBN: 9780470036341).
- [39] M. Kayri, "Predictive abilities of Bayesian regularization and Levenberg-Marquardt algorithms in artificial neural networks: A comparative empirical study on social data", *Mathem. and Comput. Appl.*, vol. 21, no. 2, 2016 (DOI: 10.3390/mca21020020).
- [40] V. S. Kushwah and G. S. Tomar, "Design and analysis of microstrip patch antennas using artificial neural network", in *Trends in Research on Microstrip Antennas*, S. Chattopadhyay, Ed. Rijeka: IntechOpen, 2017 (DOI: 10.5772/intechopen.69522).



Kanhaiya Sharma received his M.Tech. in Computer Science and Engineering from Jawahar Lal Nehru Technological University (JNTU) Hyderabad, Andhra Pradesh, India in 2011, and currently pursuing a Ph.D. in Computer Science and Engineering from the School of Technology, Pandit Deendayal Petroleum Univer-

sity since 2017. He worked as an Assistant Professor at Sri Sarathi Institute of Engineering and Technology, Nuzvid, Krishna District, Andhra Pradesh, India from October 2006 to May 2010, and at Siddhant group of institution, Maharashtra, India from 2010 to 2017. His current research interests include design and development of low-cost solutions for wireless applications, machine learning, wireless communications, microstrip antenna design, filter design, and artificial intelligence.

E-mail: sharmakanhaiya@gmail.com

Department of Computer Science and Engineering

School of Technology

Pandit Deendayal Petroleum University

Gandhinagar, Gujarat 382007, India



Ganga Prasad Pandey received his B.Tech. in Electronics and Communication Engineering from KNIT, Sultanpur, UP, an M.E. from Delhi College of Engineering (now Delhi Technological University), Delhi and a Ph.D. from Uttarakhand Technical University, Dehradun. He worked as an Assistant Professor at the

Maharaja Agasen Institute of Technology, Delhi, from 2002 to 2016. He joined Pandit Deendayal Petroleum University, Gandhinagar in 2016, working at the ICT Department. His current research interests include machine learning in antennas, energy harvesting, ME-dipoles, active, reconfigurable, frequency-agile microstrip antennas, as well as microwave/millimeter wave integrated circuits and devices.

E-mail: Gangaprasad.Pandey@sot.pdpu.ac.in

Department of Information and Communication

Technology

School of Technology

Pandit Deendayal Petroleum University

Gandhinagar, Gujarat 382007, India

Transmission of Disaster Warnings via Control Channels in Cellular Networks

Mohammed I. Aal-Nouman¹, Osamah Abdullah², Omar Hussein Salman³,
and Noor Qusay A. Al-Shaikhli¹

¹ College of Information Engineering, Al-Nahrain University, Baghdad, Iraq

² Department of Electrical Power Engineering Techniques, Al-Ma'moon University College, Baghdad, Iraq

³ Engineering College/Network Department, Al Iraqia University, Baghdad, Iraq

<https://doi.org/10.26636/jtit.2020.140820>

Abstract—According to United Nations reports, natural disasters caused, worldwide, approximately 100,000 deaths and affected 175 million people each year between 2004 and 2013. To reduce those numbers, countries around the globe have made specific arrangements enabling them to warn the population about imminent disasters, in order to evacuate the area in due time. But providing such warnings in areas where no Internet access is available poses a great challenge. In this paper, we proposed a method to transmit early warning messages via UMTS cellular networks, while relying on spare extensions of control channels (FACH). The results obtained are validate based on their comparison with theoretical considerations and are also benchmarked against the 3GPP standard. The results show that messages may be sent faster than with the use of the 3GPP standard.

Keywords—control channel, disaster reduction, early warning message, mobile network.

1. Introduction

According to reports drawn up by the United Nations (UN), more than 6,873 natural disasters were recorded worldwide between 1994 and 2013, with more than 68,000 lives lost each year, and with the average number of injuries equaling approximately 180 million. 218 million people were affected by such geophysical disasters as, volcanic eruptions, landslides, tsunamis and earthquakes [1].

Humans become increasingly vulnerable to natural disasters due to a variety of factors, such as environmental changes and population growth. Countries around the world have recognized the importance of disaster warnings systems, being a part of their risk management schemes [2]. Accurate weather forecasts and early warning systems are an integral part of disaster prevention solutions which help reduce the losses suffered [1], [3]. However, in some cases, especially in the rural areas, forecasts cannot be reliable if provided for periods over 12 hours [4].

Specific attempts have been made around the world to define adopt an approach that allows to prevent or limit the

aftermath of natural disasters and their impact on human's life. "2030 agenda for sustainable development" adopted in New York in 2015 may serve as a good example here. The meeting at which it was adopted redefined the importance of local level activities and determined goals that need to be achieved in order to make cities inclusive, safe, resilient and sustainable by the year 2030 [5]. Another meeting, held in Paris in 2015 and devoted to climate change, suggested that third-party stakeholders are welcome to participate in addressing the issue and helping reduce the outcomes of natural disasters [5].

The "Sendai framework for disaster risk reduction 2015–2030", first set up in Japan, was adopted by the UN in 2015 at the "Third UN World Conference on Disaster Risk Reduction". It is the first major agreement focusing on the 2030 development agenda and aims to reduce the risk of disasters and to limit losses in human lives. The agenda has set many targets to be achieved by 2030, such as Target G, stating that information warning about disasters should reach all people, regardless of their location [6], [7]. Such an early warning will reduce the outcomes of disasters, provided that the population is evacuated after receiving such an early alert [8].

The messages sent out may be useful even after the occurrence of a hazardous event, providing after disaster support, e.g. by making blood donation requests [8], [9] and shelter locations [10]. The importance of early warning systems may be even greater if they are combined with IT solutions [11]. However, the Internet connection may not be available or reliable outside of cities, which means that other solutions are required [12].

The Internet connection may be overloaded in areas where large numbers of users are present, e.g. in stadiums or concert halls. The speed of the connection and its availability in rural areas are worse than in large cities [13]–[15]. Therefore, importance of early warning messages is even greater in areas where the quality of Internet connections is poor. According to Human Rights Watch, Internet access is blocked in some countries during crisis situations or rev-

olutions, e.g. in Bangladesh, the Democratic Republic of Congo, Egypt, India, Indonesia, Iran, Iraq, Sudan, Myanmar and Zimbabwe [16]. This makes non-Internet solutions even more important.

According to OpenSignal and the GSM Alliance (GSMA), i.e. companies providing statistical data concerning mobile coverage, 3G mobile coverage is better compared with 2G. A study shows that in 95 countries participating in the survey, 93 offered better availability of 3G signal [17], [18]. 3G networks are available worldwide and deployed by 314 operators in 118 countries [19]. While there are some countries and operators planning to switch off 3G before 2G, other states will do the opposite and will shut down their 2G networks before 3G. Some of them have already shut down their 2G networks, e.g. Japan [20], Indonesia [21], [22], Australia [23], South Korea [24], Singapore [25], Taiwan and Thailand [18]. In some countries, operators' policies differ—some of the service providers have already switched off their 2G networks, e.g. AT&T and Verizon in the USA, Telus and Bell in Canada or Swisscom in Switzerland [26]–[29].

The majority of Central and Northern African countries, as well as South-West Asia states are still using 3G and 2G [30], and some of their operators will terminate 2G before 3G, e.g. STC in Saudi Arabia [31] and Vodacom in South Africa [32]. Other countries will follow this trend and will shut down their 2G networks, keeping 3G operational—e.g. the Netherlands [33].

There is no solution that fits all generations of cellular network solutions and that is able to send messages associated with natural disasters. This stems from the fact that the technologies used differ in so many aspects. In this paper, as an alternative to the common SMS system, we propose another method to ensure delivery of messages, because as long as the network works fine, signaling messages such as RACH and FACH will be able to deliver the alerts, even in rural areas, where the Internet connection is poor. The proposed system is using the spare extension of the FACH control channel, transmitted over the secondary control physical channel (S-CCPCH) [34]. The main contribution of this research project is to provide a method, other than an SMS-based solution, to send disaster warning messages without relying on the Internet.

In UMTS, FACH relies on the third general partnership project (3GPP) standard being a part of the radio resource control (RRC) procedure. The random-access channel (RACH) is used as well. Those two solutions are used together to connect the user's equipment (UE) to the core network [34]. Once RRC connection has been completed, other protocols and procedures commence the transmission of data. In the proposed method, data reaches the user immediately after the RRC connection has been established, without a need for completing another step, as FACH spare extension is used to transfer the warning messages. The rest of the paper is organized as follows. Similar applications and disaster message delivering methods are presented in Section 2. Section 3 provides a brief description of the

channel used in the proposed method. The procedure relied upon by the proposed approach is shown in Section 4. Section 5 analyses the model of the channel. System evaluation and system validation are described in Sections 5, 6 and 7, respectively. Finally, conclusions are presented in Section 8.

2. Related Works

Disaster warning messages have been transmitted based on the digital audio broadcasting (DAB) technology by Hongsheng [35]. DAB was relied upon for a two-way transmission of information (uplink and downlink) within the threatened areas. The messages are transmitted from the transmitter to the receiver by dongles attached to a computer operating at each end. mKRISHI is another warning system that alerts fishermen at sea or on the coast. It uses VHS transmission as the primary communication method. It has a Web-based database that stores the data and then transmits them to individual users [36].

Other researchers use sensors to check the weather conditions and then transmit information using a wireless sensor network (WSN). However, the proposed method uses GPRS to send alerts to the database in the cloud, and then a real-time video stream is sent to the users using the Internet [36]. Another approach uses WSN to collect weather-related information. It is collected by the sink node and then sent to the data center via the Internet, providing the user with disaster information based on a GPRS connection [37]. Similarly, emergency information is sent using a WSN with an ultrasound sensor, to the main system, via the Internet in [38].

In [39], Solmaz proposed a solution where mobile phones are used as sensors to warn people about an upcoming disaster. The sensor nodes (mobile phones) receive messages from the host (sink) which broadcasts the message.

Social media, such as Twitter, also could be useful in warning the population. Many researchers used them transmit alerts, but the Internet still remains the only medium that is used [39], [40]. Other social media and communicators, like WhatsApp, Facebook and Viber, may also be relied upon to request help [41].

3. FACH and S-CCPCH

UMTS-based radio access networks (UTRAN) usually send information and control data to the user using the FACH control channel, transmitting over S-CCPCH. Each S-CCPCH radio frame needs 10 ms to finish and has 15 slots, each containing 2560 chips. The slot size may be $20 \cdot 2^k$ bits, where k could be any integer number from 0 to 6, depending on the spread factor which equals 2^x (x may equal 2 to 8). Figure 1 shows the structure of S-CCPCH [34], while Table 1 gives the characteristics of each slot type.

The figure shows that size of TFCI field may be 0, 2, 4 or 8 bits, while N_{pilot} field may be 0, 8 or 16 bits. The

remaining fields will be used to transmit control-related information [34].

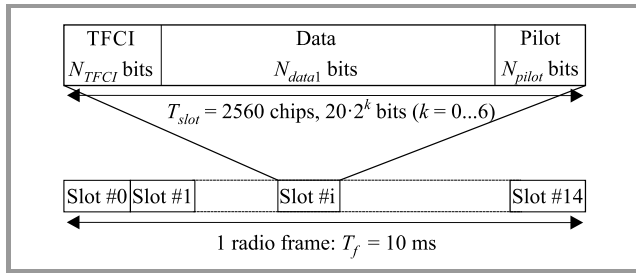


Fig. 1. S-CCPCH radio frame structure.

Table 1
Characteristics of S-CCPCH

SF	Format	Channel bit rate [kb/s]	Channel symbol rate [ks/s]	Bits per slot	Bits per frame
256	0	30	15	20	300
128	4	60	30	40	600
64	8	120	60	80	1200
32	11	240	120	160	2400
16	12	480	240	320	4800
8	15	960	480	640	9600
4	16	1920	960	1280	19200

FACH data will be filled in the slots and then mapped to S-CCPCH fields of the UMTS. If FACH information cannot be fitted into a single slot, the next slot will be used, and then if all slots of the first frame are fully loaded, the next frame will be used.

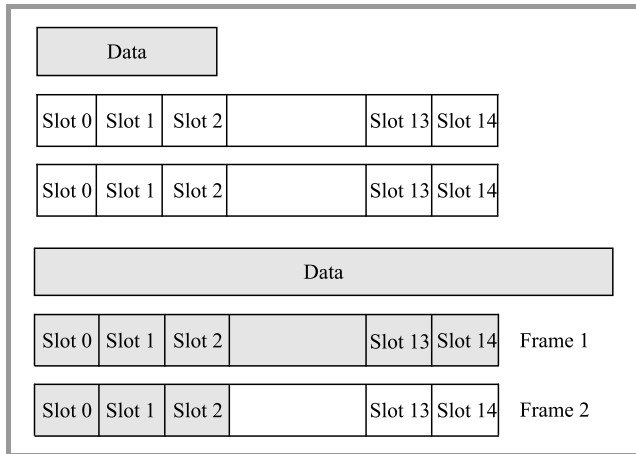


Fig. 2. Mapping data into frames and slots.

4. Proposed Method

The alerts sent may be referred to as downlink, since the warnings are from the mobile network to the UE using UMTS. The universal terrestrial radio access network UTRAN receives the alerts from the operator and a third-

party data and service provider as a part of location-based services [42]. The downlink stage comprises several, as shown in Fig. 3.

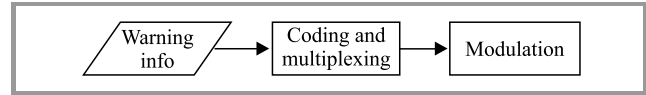


Fig. 3. Downlink steps.

FACH frame may be split into the header and the payload. The header contains such information as field type (FT) and CFN, used to indicate which frame is sent first, and showing the power level (Fig. 4).

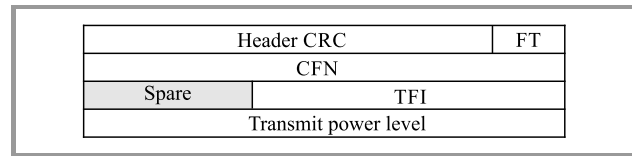


Fig. 4. FACH's header.

The size of the payload is not fixed and may change depending on specific factors, such as spreading and slot format. A change frame size may also be introduced, e.g. the first bit in the new IE FI if it is set, the AOA IE field will be presented and if it is reset the AOA will not be used. In the same way, if the second bit of the new IE FI is active, the LCR ID IE cell portion will be presented [34]. Figure 5 show the payload fields in FACH.

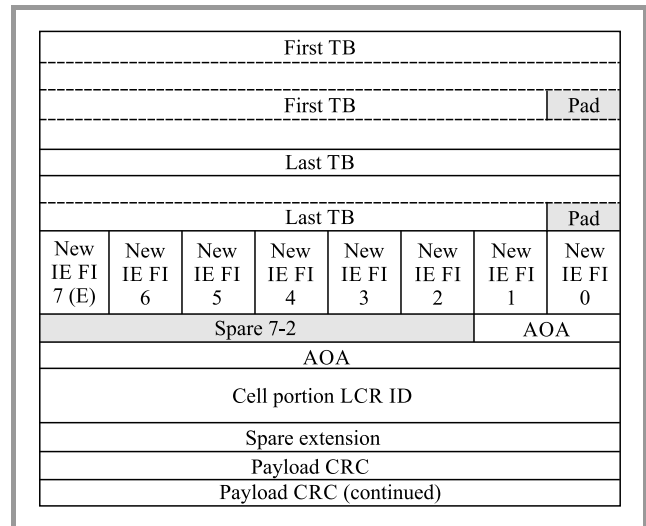


Fig. 5. FACH's payload.

Generally, spare extension fields are filled with zeros, are then transmitted, and the receiver will neglect that data [34]. In the proposed method, the spare extension of FACH is used to transmit the disaster-related message. The FACH frame is then loaded to S-CCPCH. If one slot is not capable of carrying the entire frame, the next slot will be used. Furthermore, if the data cannot be accommodated in one frame, the next frame will be utilized.

The spare extension is 28 8-bits characters in size. If the size of the information needs to be sent is lower, the remaining space of the extension will be filled zeros to reach 28. Yet, if the data size exceeds 28 characters, the second frame will be used. At the end, the CRC sum of the header is calculated and inserted into the header. The CRC value of the payload is also calculated but with a different polynomial to achieve better reliability [34]. Finally, each two-bit piece of data will be presented as one symbol and will be sent using QPSK modulation. Figure 6 shows a flowchart of the downlink process.

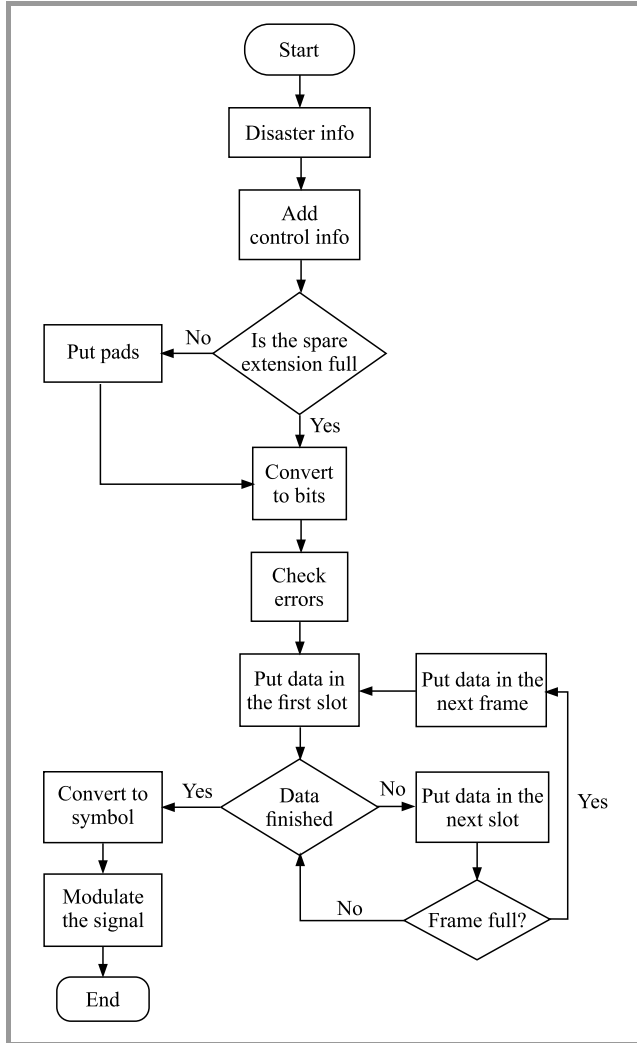


Fig. 6. Sending procedure flowchart.

On the receiver side, the UE extracts the information from the spare extension. The spreading factors of S-CCPCH may assume different values, i.e. 4, 8, 16, 32, 64, 128, and 256.

5. The Model

In this section, the analytical model of the proposed method is presented. First, the S-CCPCH calculation is shown,

which is the same for the proposed method and the 3GPP standard. Then, calculation for FACH for 3GPP standards will be presented and, finally, the FACH model will be shown.

The disaster information is loaded into the spare FACH extension. S-CCPCH carries FACH. Let an S-CCPCH frame be called X and be divided into 15 slots ($y_1 \dots y_{15}$) [43], such as:

$$X = \{y_1, y_2, \dots, y_i\} . \quad (1)$$

Each slot has N_{TFCH} and N_{pilot} and it will be set to 8 as it is the most common slot format [43]. That will leave the size of the slot as:

$$Size_y = 20 \cdot 2^k - N_{TFCH} + N_{pilot} , \quad (2)$$

$$Size_y = 20 \cdot 2^k - 8 , \quad (3)$$

where the k value could be from 0 to 6, depending on the spreading factor SF which could be 4, 8, 16, 32, 64, 128, or 256. Value K may be calculated from:

$$SF = \frac{256}{2^k} \implies k = \frac{\log \frac{256}{SF}}{\log 2} . \quad (4)$$

Because the frame size is equal to the number of slots multiplied by its number (up to 15 slots), then the frame size is:

$$Size_x = \sum_{i=1}^m Size_y , \quad m \leq 15 \text{ [bits]} , \quad (5)$$

$$Size_x = \sum_{i=1}^m (20 \cdot 2^k) - 8 \text{ [bits]} , \quad (6)$$

where $K \in \{0, 1, 2, 3, 4, 5, 6\}$ and $m \leq 15$,

$$Size_x = \sum_{i=1}^m \left(20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} \right) - 8 \text{ [bits]} , \quad (7)$$

$SF \in \{4, 8, 16, 32, 64, 128, 256\}$, and $m \leq 15$.

The channel bitrate is:

$$Ch_{bitrate} = \frac{Size_x}{Time} . \quad (8)$$

Because the time needed to finish one frame is 10 ms the channel bitrate will be:

$$Ch_{bitrate} = \frac{1}{10} \cdot \sum_{i=1}^m \left(20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} \right) - 8 \text{ [Kbps]} , \quad (9)$$

for $SF \in \{4, 8, 16, 32, 64, 128, 256\}$, $m \leq 15$.

The data size in FACH needs to be calculated in order to know how many frames S-CCPCH will need in order to send the data loaded into FACH. The FACH structure has two main parts: the header and the payload, where the header size is four bytes [34], and the payload has many fields used for configuration and control. These fields are

5 bytes long. Also, the payload section has transport block fields (TB). Each of them is 8 bits long. The spare extension section will finally comprise 28 bytes. The size of the spare extension may be up to 28 bytes, and can be calculated from:

$$SE = \left\lceil \frac{DI}{28} \right\rceil \cdot 28 - DI \text{ [bytes]}, \quad (10)$$

where SE is the spare extension size and DI is the disaster information. If DI size is below 28 bytes, SE will add pad fields to empty places. Because the information will be sent in SE , the TB field in FACH will be omitted and the payload size will be:

$$Payload = 5 + SE + DI \text{ [bytes]}, \quad (11)$$

$$Payload = 5 + \left\lceil \frac{DI}{28} \right\rceil \cdot 28 \text{ [bytes]}, \quad (12)$$

where “5” is the number of bytes for the control fields. The number of frames needed without the header can be calculated as:

$$N_{frame} = \left\lceil \frac{Payload}{Size_x - 32} \right\rceil, \quad (13)$$

$$Header_{size} = 4 \cdot N_{frame} \text{ [bytes]}, \quad (14)$$

$$Header_{size} = 8 \cdot 4 \cdot \left\lceil \frac{8 \cdot (5 + \left\lceil \frac{DI}{28} \right\rceil \cdot 28)}{\sum_{i=1}^m (20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} - 8) - 32} \right\rceil, \quad (15)$$

$$Data_{size} = Header_{size} + Payload, \quad (16)$$

$$Data_{size} = 32 \cdot \left\lceil \frac{8 \cdot (5 + \left\lceil \frac{DI}{28} \right\rceil \cdot 28)}{\sum_{i=1}^m (20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} - 8) - 32} \right\rceil + 8 \cdot \left(5 + \left\lceil \frac{DI}{28} \right\rceil \cdot 28 \right). \quad (17)$$

The number of slots needed will be:

$$N_{slots} = \left\lceil \frac{Data_{size}}{Size_y} \right\rceil, \quad (18)$$

$$N_{slots} = \left\lceil \frac{Data_{size}}{20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} - 8} \right\rceil. \quad (19)$$

The time required to transmit these slots will be equal to the number of slots multiplied by 10/15 ms:

$$Time = \frac{10}{15} \cdot \left\lceil \frac{Data_{size}}{20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} - 8} \right\rceil. \quad (20)$$

Then, the number of frames will be:

$$N_{frame2} = \left\lceil \frac{N_{slots}}{15} \right\rceil. \quad (21)$$

Similarly, because the 3GPP standard uses the spare extensions as pads, the number of frames needed for the 3GPP standard is:

$$N_{frame} = \left\lceil \frac{32 \cdot \left\lceil \frac{8 \cdot (33 + TB)}{\sum_{i=1}^m (20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} - 8) - 32} \right\rceil + 8 \cdot (33 + TB)}{15 \cdot (20 \cdot 2^{\frac{\log \frac{256}{SF}}{\log 2}} - 8)} \right\rceil. \quad (22)$$

6. System Validation

Here, results of the analytical calculation are compared with the simulation. The results show SF most frequently used

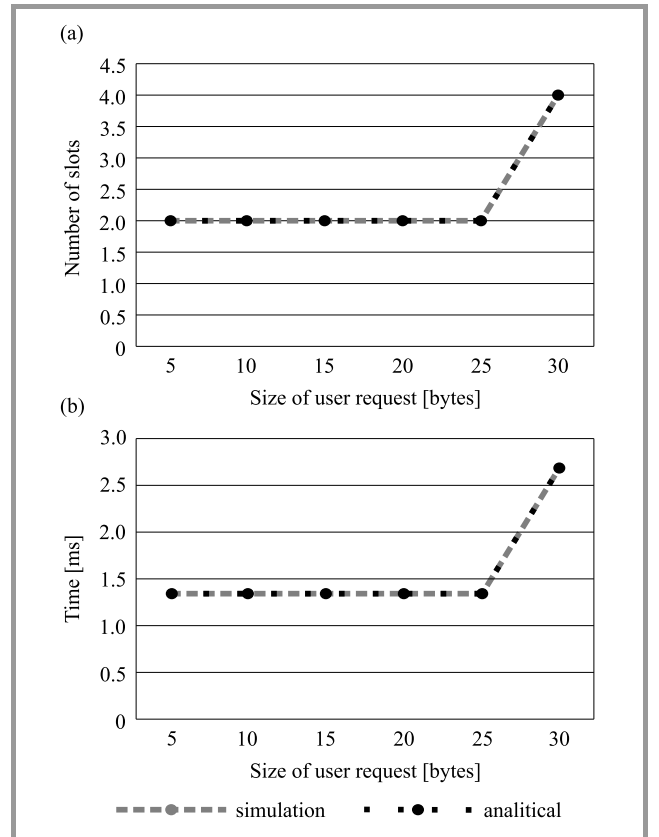


Fig. 7. Comparison of results with $SF = 32$ for: (a) slot needed, (b) time needed to transmit the info.

in the channel, which equal 32 or 64. Figure 7 shows the comparison with $SF = 32$.

Figure 7a shows the number of slots needed in S-CCPCH to send data to the user, while Fig. 7b shows the time that these slots need in order to be sent to the user. The results indicate a slight difference between the calculation results and the results obtained from the simulation. This difference could be bigger when using a SF larger than 32, as can be seen in Fig. 8, where $SF = 64$.

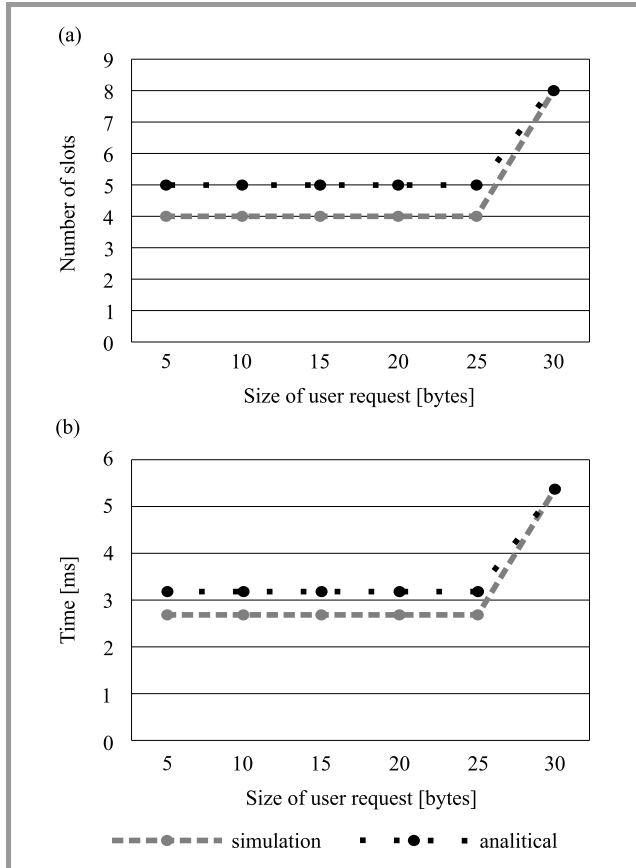


Fig. 8. Theoretical vs simulation results with $SF = 64$ for: (a) slots needed for the transmission and (b) time needed to transmit the alert.

7. System Evaluation

This section shows the results of implementing the proposed method under a simulation scenario and compares them with the 3GPP standard. The results shown are for $SF = 32$ and 64. Figure 9 presents the results of the two methods (proposed vs. 3GPP) with $SF = 32$, indicating the differences in the number of slots needed to send the information, and in the time needed to send those slots over S-CCPCH.

The performance of the proposed method is better than that of 3GPP in terms of speed at which an alert of the same size is sent. The proposed method needs less data and uses fewer slots. The reason is that the spare extensions in FACH are filled by zeros by the sender, and, TB

fields will be used to send the message, meaning that the transmission will consist of the disaster-related information plus the size of the spare extension, while in the proposed method the spare extension will be used to send the information, and the TB field will not carry any data. This will reduce the amount of data sent. Figure 10 shows the result for $SF = 64$.

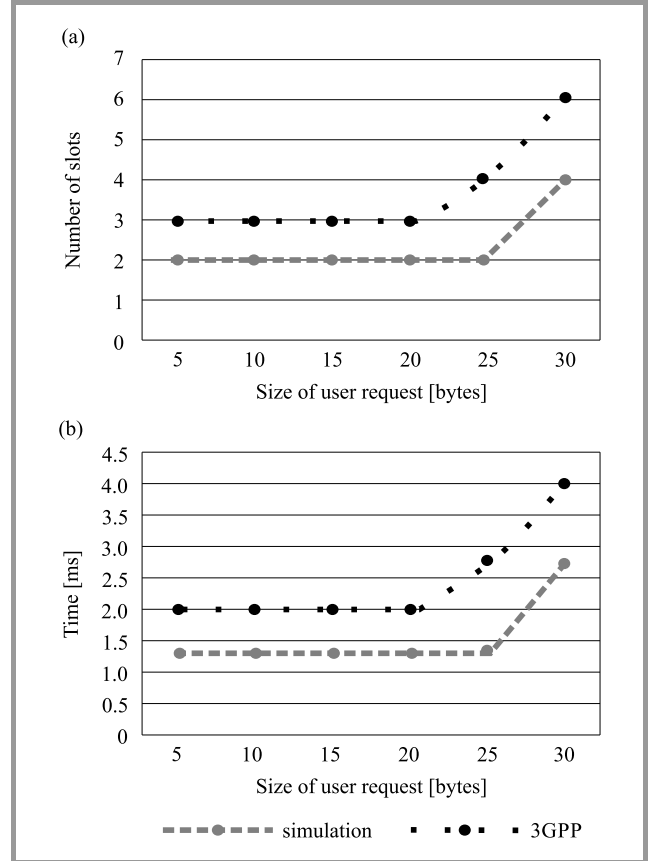


Fig. 9. 3GPP vs proposed results with $SF = 32$ for: (a) slots needed for the transmission and (b) time needed to transmit the info.

While this paper focused on transmitting the message using a 3G (UMTS) network, the idea of the proposed method can also be applied with modifications in LTE and LTE-A networks. Fortunately, LTE service providers may use the circuit switched network of UMTS and GSM when needed by the circuit switch fallback (CSFB) [44]. Hence, the proposed system may be applied in LTE and LTE-A to send warning messages. However, not all operators have the privilege of using CSFB. In this case, the proposed design may also be applied, but with different frame structures and control channels.

As mentioned earlier, UMTS uses RACH/FACH to establish a radio resource control (RRC) connection [34]. However, the channels that are used in LTE for this process, as part of RRC, are not the same as in UMTS. LTE uses PRACH and RACH with frame formats that differ from that used in UMTS. Different physical channels are used

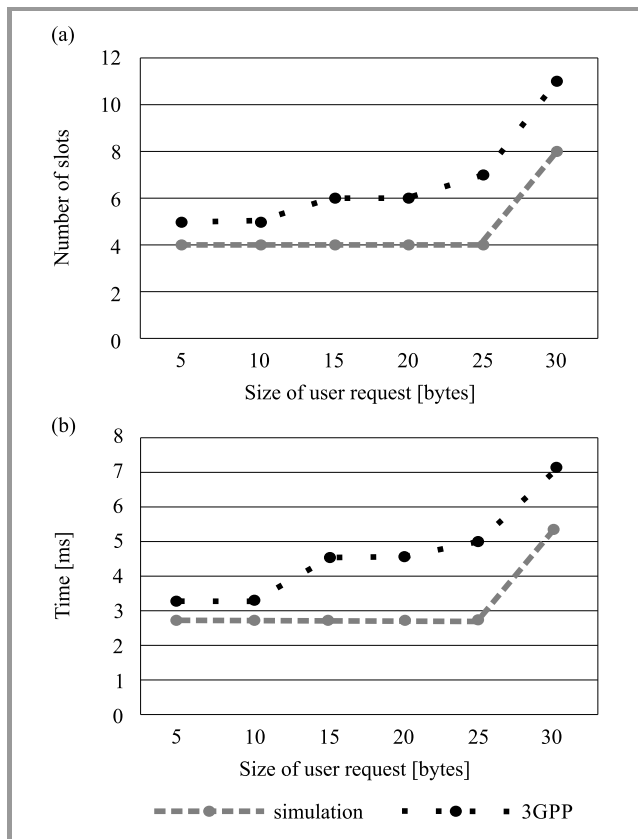


Fig. 10. 3GPP vs proposed method with $SF = 64$ for: (a) slots needed to transmit and (b) time needed to transmit the info.

as well. So, the proposed design may be applied in LTE, but with modifications in frame structure and channels used [45].

8. Conclusion

This paper proposed a new method for sending alert messages from the service provider to mobile users without the need of using the Internet or packed channels. The spare FACH extensions are used to carry the messages in UMTS. FACH is being carried using S-CCPCH.

The main contribution of this paper lies in the utilization of the spare extensions of FACH for transmitting information to the user. The spare extensions are not used in the standards as padding or guards, as they are related to FACH, not S-CCPCH.

In the 3GPP standards, FACH is mainly a part of the RRC used to establish the connection, with subsequent commencement of the process of transmitting information between the user and the mobile network. However, in many cases, FACH could carry a small amount of data using TB fields, and the transmission is the fastest in this case. So, all results shown in this paper have been compared with this scenario. Results of the evaluation show that the proposed method is capable of transmitting warning messages to users. Moreover, this transmission is faster than that in 3GPP.

References

- [1] "The human cost of natural disasters: A global perspective", *UN Office for Dis. Risk Red., Centre for Research on the Epidemiology of Dis. (CRED)* 2015 [Online]. Available: <https://climate-adapt.eea.europa.eu/metadata/publications/the-human-cost-of-natural-disasters-2015-a-global-perspective/11258275>
- [2] C. B. Mayhorn and A. C. McLaughlin, "Warning the world of extreme events: A global perspective on risk communication for natural and technological disaster", *Safety Sci.*, vol. 61, pp. 43–50, 2014 (DOI: 10.1016/j.ssci.2012.04.014).
- [3] M. Azam, H. S. Kim, and S. J. Maeng, "Development of flood alert application in Mushim stream watershed Korea", *Int. J. of Dis. Risk Red.*, vol. 21, pp. 11–26, 2017 (DOI: 10.1016/j.ijdr.2016.11.008).
- [4] C. Roy, S. K. Sarkar, J. Åberg, and R. Kovordanyi, "The current cyclone early warning system in Bangladesh: providers' and receivers' views", *Int. J. of Dis. Risk Red.*, vol. 12, pp. 285–299, 2015 (DOI: 10.1016/j.ijdr.2015.02.004).
- [5] E. A. Gencer, "Local government powers for disaster risk reduction: A study on local-level authority and capacity for resilience", *UN Office for Dis. Risk Red., Center for Urban Dis. Risk Red. & Resilience*, 2017 [Online]. Available: https://www.unisdr.org/campaign/resilientcities/assets/toolkit/documents/LG%20Powers%20for%20DRR/_2017_Final/_20170531.pdf
- [6] "The Sendai framework for disaster risk reduction 2015–2030", *UN Office for Dis. Risk Red.*, 2015 [Online]. Available: https://preventionweb.net/files/43291_sendaiframeworkfordrren.pdf
- [7] "How to make cities more resilient: a handbook for local government leaders: a contribution to the global campaign 2010–2015: making cities resilient – my city is getting ready!", *UN Office for Dis. Risk Red.* Geneva, Switzerland, 2012 [Online]. Available: https://preventionweb.net/files/26462_handbookfinalonlineversion.pdf
- [8] H. Nakamura, H. Umeki, and T. Kato, "Importance of communication and knowledge of disasters in community-based disaster-prevention meetings", *Safety Sci.*, 2016 (DOI: 10.1016/j.ssci.2016.08.024).
- [9] N. Kapucu, "Disaster and emergency management systems in urban areas", *Cities*, vol. 29, no. S1, pp. S41–S49, 2012 (DOI: 10.1016/j.cities.2011.11.009).
- [10] M. Basu, S. Ghosh, A. Jana, S. Bandyopadhyay, and R. Singh, "Resource mapping during a natural disaster: A case study on the 2015 Nepal earthquake", *Int. J. of Dis. Risk Red.*, vol. 24, pp. 24–31, 2017 (DOI: 10.1016/j.ijdr.2017.05.020).
- [11] B. Balis, T. Bartynski, M. Bubak, G. Dyk, T. Gubala, and M. Kasztelnik, "A development and execution environment for early warning systems for natural disasters", in *13th IEEE/ACM Int. Symp. on Cluster, Cloud and Grip Computing*, Delft, Netherlands, 2013 (DOI: 10.1109/CCGrid.2013.101).
- [12] Z. N. K. Wafi, M. F. A. Malek, S. H. Alnajjar, and R. Badlishah Ahmad, "Early warning system for disaster management in rural area", in *2015 Int. Symp. on Tech. Management and Emerging Technol. (ISTMET)*, Langkawai Island, Malaysia, 2015 (DOI: 10.1109/ISTMET.2015.7359061).
- [13] J. M. Cole and B. L. Murphy, "Rural hazard risk communication and public education: Strategic and tactical best practices", *Int. J. of Disaster Risk Red.*, vol. 10, Part A, pp. 292–304, 2014 (DOI: 10.1016/j.ijdr.2014.10.001).
- [14] R. Ramirez, "A model for rural and remote information and communication technologies: a Canadian exploration", *Telecommun. Policy*, vol. 25, no. 5, pp. 315–330, 2001 (DOI: 10.1016/S0308-5961(01)00007-6).
- [15] B. L. Murphy, "Emergency management and the August 14th, 2003 Blackout", *ICLR Research*, 2004 (ISBN: 0973379561).
- [16] "Shutting down the internet to shut up critics" *Human Right Watch*, 2020 [Online]. Available: <https://hrw.org/world-report/2020/country-chapters/global-5>

- [17] "Global state of mobile networks", *Opensignal*, 2020 [Online]. Available: <https://opensignal.com/reports/2016/08/global-state-of-the-mobile-network>
- [18] "World Time Zone", [Online]. Available: <https://worldtimezone.com/4g.html>
- [19] "2G, 3G, 4G... OMG! What G is Right for M2M?", *Aeris Communications* 2013 [Online]. Available: https://connectedworld.com/wp-content/uploads/2014/07/Whitepaper_Aeris_2G3G4GOMG.pdf
- [20] V. Parameswaran and M. Shanmugasundaram, "Journey of wireless communication", in *Lecture Notes in Networks and Systems*, G. Ranganathan, J. Chen, Á. Rocha, Eds. *Inventive Commun. and Computational Technol.*, vol. 89, Springer, Singapore, 2020 (DOI: 10.1007/978-981-15-0146-3_28).
- [21] T. Hidayat, R. Mahardiko, and M. Alaydrus, "Mobile cellular technology forecast for the Indonesian telecommunications industry", *J. of Telecommun. and the Digital Economy*, vol. 8, no. 1, pp. 37–48, 2020 (DOI: 10.18080/jtde.v8n1.226).
- [22] A. S. Arifin and A. Fadrian, "Study on sustainability of Indonesia 2G operator", in *2019 16th Int. Conf. on Quality in Research (QIR): Int. Sym. on Electrical and Computer Eng.*, Padang, Indonesia, pp. 1–5, 2019 (DOI: 10.1109/QIR.2019.8897909).
- [23] "Vodafone to switch off 2G network in 2018", *Vodafone*, 2016 [Online]. Available: <https://vodafone.com.au/media/vodafone-to-switch-off-2g-network-next-year>
- [24] N. Kwon and Y. Kwon, "Terminating 2G service in Korea: Policy issues and suggestions", *Int. Tel. Policy Review*, vol. 21, no. 1, pp. 1–26, 2014 [Online]. Available: <https://ssrn.com/abstract=2418848>
- [25] "Infocomm media development authority, 2G services to cease on 1 April 2017", 2017 [Online]. Available: <https://imda.gov.sg/news-and-events/Media-Room/Media-Releases/2017/2g-services-to-cease-on-1-april-2017>
- [26] "The Sunset of 2G/3G Is Coming. Is Your Business Ready?", *James Brehm Associates*, 2016 [Online]. Available: <https://business.att.com/content/dam/attbusiness/briefs/are-you-ready-for-2G-3G-sunset-white-paper.pdf>
- [27] "AT&T has officially shut down its 2G network", 2017 [Online]. Available: https://www.gsmarena.com/at_t_has_officially_shut_down_its_2g_network-blog-22811.php
- [28] *Mobile World Live*, [Online]. Available: <https://mobile-worldlive.com>
- [29] "Bye bye 2G", *Swisscom*, [Online]. Available: <https://swisscom.ch/en/about/company/portrait/network/5g/2g-phase-out.html>
- [30] P. Á. Colás, J. Bada, and J. Guerrero "The status of the 2G/3G network sunset", 2019 [Online]. Available: <https://nae.global/en/the-status-of-the-2g-3g-network-sunset>
- [31] "STC confirms switching-off its 3G network", 2020 [Online]. Available: <https://www.stc.com.sa/wps/wcm/connect/english/stc/NewsItems/stc-confirms-switching-off-3g-network>
- [32] MyBroadband, "Vodacom wants to turn off its 2G service in South Africa", *Business Tech*, 2019 [Online]. Available: <https://businesstech.co.za/news/telecommunications/334999/vodacom-wants-to-turn-off-its-2g-service-in-south-africa/>
- [33] T-Mobile Netherlands, "T-Mobile schakelt het 2G-netwerk in 2020 uit" [Online]. Available: <https://www.t-mobile.nl/2g-uitzetten> [in Dutch]
- [34] "UTRAN Iub interface user plane protocols for Common Transport Channel data streams", *3GPP*, vol. 14, 2017 [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1207>
- [35] H. Zang, G. Wang, M. Lu, D. Wang, and P. Xu, "Emergency warning and bidirectional communication via digital audio broadcast", in *IEEE Int. Conf. on Consumer Electronics (ICCE)*, Las Vegas, USA, 2017 (DOI: 10.1109/ICCE.2017.7889248).
- [36] D. Singh, D. Piplani, S. Shinde, and S. Karthik, "mKRISHI fisheries: a case study on early warning system (EWS) for disaster communication and management", in *2016 IEEE Int. Symp. on Technol. and Society (ISTAS)*, pp. 1–6, Kollam, India, 2016 (DOI: 10.1109/ISTAS.2016.7764280).
- [37] Z. Chen, C. Fang, and R. Deng, "Research and application of Jing-gangshan geological disaster prevention system based on wireless sensor network system", in *2015 23rd Int. Conf. on Geoinformatics*, Wuhan, China, 2015 (DOI: 10.1109/GEOINFORMATICS.2015.7378716).
- [38] M. Erd, F. Schaeffer, M. Kostic, and L. M. Reindl, "Event monitoring in emergency scenarios using energy efficient wireless sensor nodes for the disaster information management", *Int. J. of Disaster Risk Red.*, vol. 16, pp. 33–42, 2016 (DOI: 10.1016/j.ijdr.2016.01.001).
- [39] G. Solmaz and D. Turgut, "Modeling pedestrian mobility in disaster areas", *Pervasive and Mobile Comput.*, vol. 40, pp. 104–122, 2017 (DOI: 10.1016/j.pmcj.2017.05.005).
- [40] P. M. Landwehr, W. Wei, M. Kowalchuck, and K. M. Carley, "Using tweets to support disaster planning, warning and response", *Safety Science*, vol. 90, pp. 33–47, 2016 (DOI: 10.1016/j.ssci.2016.04.012).
- [41] K. M. Carley, M. Malik, P. M. Landwehr, J. Pfeffer, and M. Kowalchuck, "Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia", *Safety Science*, vol. 90, pp. 48–61, 2016 (DOI: 10.1016/j.ssci.2016.04.002).
- [42] M. Aal-Nouman, O. H. Salman, H. Takruri-Rizk, and M. Hope, "A new architecture for location-based services core network to preserve user privacy", in *2017 Annual Conf. on New Trends in Information & Commun. Technol. Applications (NTICT)*, pp. 286–291, Baghdad, 2017 (DOI: 10.1109/NTICT.2017.7976118).
- [43] "Physical channels and mapping of transport channels onto physical channels (FDD)", *3GPP*, vol. 14, 2017 [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=467>
- [44] "Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2", 2020 [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123200_123299/123272/16.00.00_60/ts_123272v160000p.pdf
- [45] "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification", *3GPP*, vol. 16, 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2440>



Mohammed Aal-Nouman received his Ph.D. degree in mobile communications from University of Salford, Manchester, UK in 2017. Since 2008, he has been working as a lecturer at Al-Nahrain University, College of Information Engineering. He teaches various subjects, including, inter alia, mobile networks, wireless networks, microprocessors and computer architectures. His research interest includes mobile and wireless communications, eHealth systems and location-based services.

 <https://orcid.org/0000-0001-5056-8567>

E-mail: m.aal-nouman@coie-nahrain.edu.iq

Department of Information and Communication Engineering

College of Information Engineering

Al-Nahrain University

Baghdad, Iraq



Osamah Abdullah received his Ph.D. in Electrical and Computer Engineering from Western Michigan University in 2016. His teaching experience and practical skills are in communication systems and signal processing. Relying on his research background, he currently focuses on machine learning algorithms, UAV channel modeling,

brainwave signals detections, mobile communication, signal processing, and peer-to-peer/ad-hoc networking, with a particular emphasis placed on group location, tracking and monitoring applications.

<https://orcid.org/0000-0002-6953-7346>

E-mail: osamah.abdullah@wmich.edu

Department of Electrical Power Engineering Techniques
Al-Ma'moon University College
Baghdad, Iraq



Omar Hussein Salman received his M.Sc. degree in Computer Engineering from Al-Nahrain University, Baghdad, Iraq and his Ph.D. in Computer Systems Engineering from University Putra Malaysia (UPM) in 2007 and 2016, respectively. In 2008 he joined the academic staff of the Al-Iraqia University. Engineering College, Net-

work Department, as a lecturer and research consultant.

His research interests include data processing, algorithms, remote applications, bioinformatics, telemedicine, IoT, big data, data mining and healthcare services.

<https://orcid.org/0000-0002-4415-1236>

E-mail: omar.salman@aliraqia.edu.iq

Network Department
Engineering College
Al Iraqia University
Baghdad, Iraq



Noor Qusay A. Al-Shaikhli obtained an M.Sc. degree in Information and Communication Engineering from Al-Nahrain University, Iraq in 2019, and a B.Sc. degrees from Greenwich University, UK and MSA University, Egypt, majoring in Electronics and Communication Engineering in 2011. Since 2013, she has been working at

Al-Nahrain University, College of Information Engineering, where she is currently an assistant lecturer. Her research interests include antenna design, simulation and fabrication, wireless communication and FPGA, mobile communications, and 5G communication.

<https://orcid.org/0000-0001-7765-174X>

E-mail: noor.qusay@coie-nahrain.edu.iq

Department of Information and Communication Engineering
College of Information Engineering
Al-Nahrain University
Baghdad, Iraq

Detection of DDoS Attacks in OpenStack-based Private Cloud Using Apache Spark

Shweta Gumaste, Narayan D. G., Sumedha Shinde, and Amit K

KLE Technological University, Hubballi, India

<https://doi.org/10.26636/jtit.2020.146120>

Abstract—Security is a critical concern for cloud service providers. Distributed denial of service (DDoS) attacks are the most frequent of all cloud security threats, and the consequences of damage caused by DDoS are very serious. Thus, the design of an efficient DDoS detection system plays an important role in monitoring suspicious activity in the cloud. Real-time detection mechanisms operating in cloud environments and relying on machine learning algorithms and distributed processing are an important research issue. In this work, we propose a real-time detection of DDoS attacks using machine learning classifiers on a distributed processing platform. We evaluate the DDoS detection mechanism in an OpenStack-based cloud testbed using the Apache Spark framework. We compare the classification performance using benchmark and real-time cloud datasets. Results of the experiments reveal that the random forest method offers better classifier accuracy. Furthermore, we demonstrate the effectiveness of the proposed distributed approach in terms of training and detection time.

Keywords—cloud, DDoS, distributed processing, OpenStack, Apache Spark, random forest.

1. Introduction

Cloud computing is a model of computing resources, such as storage devices, servers, services, applications and networks, that are accessible from any location via the Internet. Cloud technology enables hardware infrastructure to be considered a common shared service on which applications and different services are provided in a cost-effective manner.

Private cloud is one of the deployment models which provides computing services offered over a private internal network. However, security issues are the primary concern affecting the data entered by and the cloud services relied upon by users. DDoS attacks are one of the major threats that are faced in cloud computing [1], [2]. The main strategy behind a DDoS attack is to target the victim by submitting a lot of requests in a distributed manner in order to exhaust the victim's resources, and therefore rendering the target resources unavailable, over a specific period of time, to the legitimate users. Hence, detection of such attacks in

the cloud requires that special attention be paid. OpenStack is an open-source software platform that provides a set of software tools for creating and managing virtual servers. When a private cloud is deployed using OpenStack, the firewall alone is not sufficient to counter DDoS attacks, so the administrator requires an additional intrusion detection system supplementing the firewall in order to detect DDoS attacks.

Classification algorithms are utilized to classify traffic packets and to predict intrusions. Selection of the appropriate machine learning algorithm for each dataset is an important issue. For accurate classification of data, one must select the right classifier. Thus, the estimation of the algorithms and the comparison of their performance are necessary while choosing the classifier. Furthermore, there is a trade-off between choosing classifiers based on their accuracy, precision and computational time. Most classifiers suffer from issues in terms of the pace of intrusion detection and the related performance. This motivates researchers to seek new approaches enabling more efficient and faster detection of DDoS attacks.

The system we propose follows an anomaly-based detection approach in a private cloud. Network traffic from the cloud is captured and analyzed using a model trained to detect intrusions. Next, we deploy these algorithms on a distributed platform for faster processing. The Apache Spark framework is used for the detection of anomalous traffic in the data captured. Spark uses a stream-oriented approach, relying on in-memory computations to process real-time data faster. This helps in real-time detection of DDoS attacks. An efficient classification model is selected for the classification of traffic packets.

The contributions of the work are as follows:

- we designed Spark as a service in an OpenStack cloud for provisioning of on-demand Spark clusters,
- we determined the best classifier for detection of DDoS attacks,
- we designed a DDoS attack detection system in a distributed framework using Spark,

- we evaluated the proposed system using a real-time cloud testbed.

The rest of this paper is organized as follows. In Section 2, we discuss DDoS attacks, related works and the Apache Spark framework. Section 3 describes the proposed methodology and the algorithms used. Section 4 presents the results of machine learning classifiers and the DDoS detection system deployed on the experimental testbed. Finally, conclusions and future work are presented in Section 5.

2. Background Study

In this section, we introduce the concept of DDoS attacks. Next, we discuss the related research on DDoS attacks detection, focusing on cloud environment and distributed processing. Furthermore, we discuss features of Apache Spark in order to facilitate understanding of the distributed approach to make the system scalable and efficient.

2.1. Distributed Denial of Service Attacks

DDoS attacks are the most widely used type of cyber-attacks where the attacker uses a botnet consisting of malware-infected machines. As shown in Fig. 1, the attacker uses a command and control server and the botnet to perform a DDoS attack on the victim machine. The foe machine specifies the attack type and the victim's IP address to the command and control server. Then, the command and control server sends this information to the botnet network. Multiple machines comprising the botnet network send a stream of attack requests to the victim machine, by throwing a wrong input or by sending a huge number of requests simultaneously. Even after the victim server responds to this request, the botnet machines throw the same request again and again. Thus, resources of the victim server become exhausted, the server crashes and its performance degrades. The scenario in which the victim's server is bombarded with a huge number of requests leading to server crash, is called a denial of service, as authorized

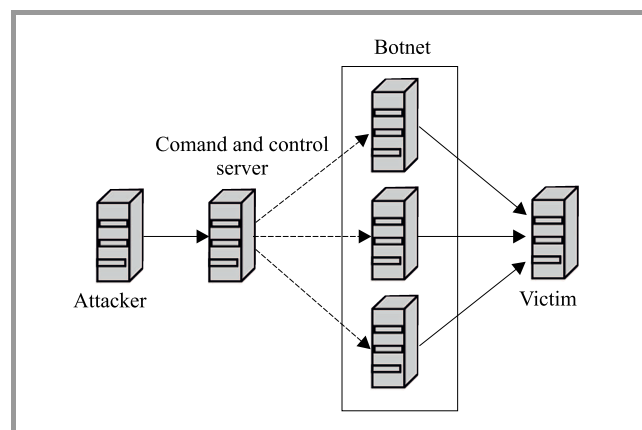


Fig. 1. DDoS attack scheme.

users are unable to access the server or a given service. As the attacks originate from different machines of the botnet, they are referred to as distributed DoS attacks. As the number of DDoS attacks has been increasing in recent years, detection or mitigation of DDoS attack has become an important issue.

2.2. Related Work

An intrusion detection system (IDS) is a security mechanism deployed for detecting malicious activities. IDS involves two types of detection methods that may be signature- or anomaly-based. Signature-based IDS uses the signature of known attacks. Snort is an open source IDS that relies on signature-based techniques for detecting attacks known from experiments [3]. The authors presented a DDoS detection mechanism based on machine learning and signature detection techniques. In [4], the authors proposed an NIDS framework operating in the cloud and consisting of a snort and signature-based apriori algorithm to detect known attacks in the network. However, signature-based mechanisms cannot detect unknown attacks or variants of known attacks.

Anomaly-based detection captures network data or logs to detect anomalies. It detects a new form of attacks rather than their signatures only. In [5], the authors proposed an entropy-based DDoS attack detection approach deployed in a cloud-based network system. The normalized entropy value is calculated to check changes in the randomness between packet header field samples. This system is maintained and represented by a third party which sends a notification to the client. In [6], the authors proposed a DDoS detection mechanism that focuses on various features of attack packets. The features are obtained by studying incoming network traffic. The features are then analyzed using radial basis function (RBF) neural networks. The authors evaluated the proposed method using a simulated network and the UCLA dataset. In [7], the authors presented an IDS operating in the cloud environment and focusing on the software-as-a-service (SaaS) model to provide application level security to the cloud customer. They consider a virtual private network for information exchange. However, research has not been conducted with regard to real cloud infrastructures. The authors in [8] present a DDoS attack detection methods using OpenStack cloud. OpenStack is an open source cloud operating system which is used for academic research and for commercial cloud applications.

The above techniques do not use distributed processing for detecting DDoS attacks. In cloud environments, huge numbers of packets or connections with customers' networks exist. Thus, protecting the organizations' networks from DDoS attacks requires huge amounts of processing power. So, there is a trend towards using distributed environments, like Hadoop and Spark, for detection of DDoS attacks. The work proposed by [9] focuses on a reliable DDoS attack detection approach based on HTTP GET flooding. It uses MapReduce processing for flooding detection. The results

show that processing time is better than in snort detection. In [10], the authors designed a DDoS prevention scheme based on the Hadoop framework, using the MapReduce technique and a genetic algorithm (GA) to detect and prevent DDoS attacks. They generate the filtering rules with an entropy-based detection scheme. The rules are updated accordingly as well. In [11], the authors presented a DDoS attack detection system utilizing classification algorithms in the Spark framework. In this work, T-shark is used for network analysis. The fuzzy logic algorithm chooses the next candidate for classification. Classification algorithms are designed to predict DDoS attacks affecting traffic packets.

The authors of [12] proposed a DDoS detection system which uses the feature selection technique known as ensemble-based multi-filter to select optimal features by combining the output of four filter methods. The method is evaluated using an NSL-KDD dataset and a decision tree classifier. However, the discussion covers the efficient feature selection technique only. Furthermore, the system suffers from an overhead in the form of a processing delay. In [13], the authors use the Hadoop map to reduce the architecture for faster processing of log files. Furthermore, abnormal behavior of sources that generate packets erratically is predicted. The architecture uses prediction-based time series analysis, which helps in faster detection of suspicious hosts. However, research is limited to threshold-based detection.

In [14], the authors proposed a live DDoS detection mechanism to detect 4 types of flooding attacks in real-time, using the map reduce approach. The results were validated on a real-time testbed using low cost hardware. The authors of [15] proposed a technique to detect DDoS attacks using the Hadoop framework for processing and analyzing large scale attacks in real-time. The entropy of source addresses is used as a metric for detection. The results are validated using the testbed. In [16], the authors proposed a distributed and collaborative approach to the processing of a large amount of data by distributing it among a number of mappers and reducers. The proposed work is validated using benchmark datasets, as well as real datasets generated using the experimental testbed, based on various metrics.

In [17], the authors proposed a distributed and collaborative technique for early detection of DDoS attacks and flash crowds. However, due to non-cooperation between various ISPs, it is difficult to modify routers to detect attacks in real-time. The authors validate the results using a real-time testbed. In [18] the authors designed a DDOS attack detection mechanism using different classifiers for OpenStack cloud. The authors validate the results using a testbed with real-time and benchmark datasets. The authors also proposed a DDoS attack detection mechanism in the SDN environment using a Mininet simulator [19]. The authors use SVM and DNN to classify attacks by capturing real-time packets. However, the authors of [18], [19] have not used any distributed environments for the detection of DDoS attacks. As the millions of packets are cap-

tured per second in the cloud, the design of a distributed processing approach relied upon by the detection system constitutes a crucial research issue.

2.3. Distributed Processing using Spark

Apache Spark is a distributed platform built on the YARN infrastructure. Spark supports in-memory processing which is iterative in nature. Spark supports many languages, such as Scala, R, Java and Python. As Python is an efficient high-level language and allows us to develop the system faster, we use Python. The proposed system employs RDD, Spark streaming and MLlib features. RDD is the key for Spark to offer failure recovery and data dependency functionalities. RDD stores intermediate results in the memory, which significantly improves computation speed. Unlike MapReduce in Hadoop, MapReduce in Spark is well packaged into RDD.

MLlib is Spark's distributed machine learning (ML) library. It provides ML algorithms with libraries used for such purposes as regression, classification, clustering, collaborative filtering, etc. Spark streaming is used for real-time calculation that relies on RDD. RDD helps in seamless connection with Spark to fuse historical data and real-time data. Spark streaming divides the streaming data into small time intervals and into RDD data sets. Later, it processes the RDD in batches and helps process complex streaming data faster. Thus, RDD helps in processing huge numbers of packets in real-time, extracts the features and detects the attacks affecting cloud environments.

3. Proposed Methodology

This section offers a description of the proposed system that allows to detect DDOS attacks using the distributed processing approach. The chapter has been divided into two subsections, focusing on the design of a Spark cluster as a service on an OpenStack private cloud, and on the detection of DDoS attacks using real-time cloud traffic and the KDD Cup dataset, respectively.

3.1. Spark as a Service

The design of this service involves setting up an OpenStack private cloud and configuring a multi-node setup. The multi-node setup is done on virtual machines (VMs) by configuring compute, neutron and controller nodes. The VMs are spawned on the compute node of OpenStack. Deployment of the Spark cluster is done, firstly, by executing the machines with single node Spark configuration steps. One of the single nodes is designated as a master and the remaining nodes are configured to be working as slaves under the supervision of the master. The user interface is provided by relying on the Django framework with Python. The user can select the number of slaves required. The user is provided with options to select the number of VC-PU's, while RAM and disk size are required for the master

and the slaves based on standard templates. The user can also create his/her own flavors for slaves. Once the user clicks the launch button, the Spark cluster will be installed in the cloud with the help of an automation script executed at the OpenStack backend using controller and compute nodes. The design includes a connection between Spark and OpenStack, as shown in Fig. 2.

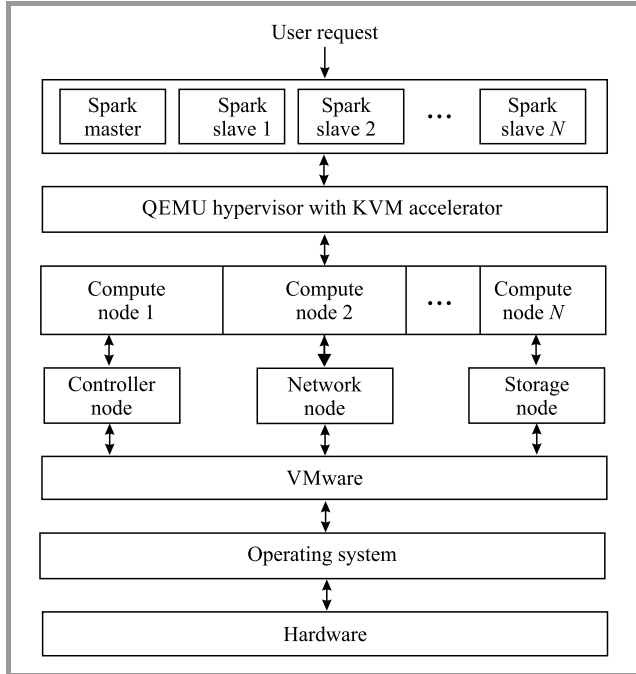


Fig. 2. Spark as a service.

The algorithm for automation of Spark as a service is described in Algorithm 1. Spark is built using Hadoop architecture. Hence, it involves Hadoop installation. Hadoop uses a Java framework. For big data processing MapReduce is required, which is written in Java. Java is used for analyzing and processing large data sets. So, JDK is installed for supporting the Hadoop platform.

3.2. DDoS Detection System

The incoming packets get captured through a packet sniffer stored in the local disk as packet log files in CSV format. In real life, these log files need faster processing. So, a distributed computing platform is used for big data analysis and for faster detection of attacks. Anomaly-based IDS treats any deviation from the normal pattern as an attack. As shown in Fig. 3, the model of the proposed system consists mainly of four modules, i.e. packet sniffer, pre-processing using Spark steaming, classification, and detection. All of them are discussed in detail below.

3.3. Packet Sniffer

Packet sniffer is designed using socket programming to capture and analyze network traffic by connecting two virtual machines on a cloud network in order to enable them to

Algorithm 1. Spark as a service on OpenStack cloud

Input: MapReduce cluster name C , number of slave nodes N , configuration of master and slave nodes S , keypair name K

Output: Cluster with required number of nodes

- 1: Create VMs based on master and slave size S using OpenStack SDK
- 2: Create a keypair K
- 3: Create one node as master node
- 4: Create N slave nodes with configuration S
- 5: Assign IP address for the master node
- 6: Change the hostname in hosts file and hostname file
- 7: **For** N node
- 8: Assign the IP address to slave nodes
- 9: Change the host name in hosts and hostname files
- 10: **End for**
- 11: Add all the IPs of master and slave nodes in private IP file
- 12: Install the services on master and slave nodes
- 13: Configure SSH on master node
- 14: **For** N nodes
- 15: Configure SSH on slave node
- 16: **End for**
- 17: Send the keypair from master to slave nodes
- 18: Disable the password authentication
- 19: Do the Hadoop name node format on master node
- 20: Start all the services on master node
- 21: Write all the IPs to one file
- 22: Mail the keypair and master node IP address

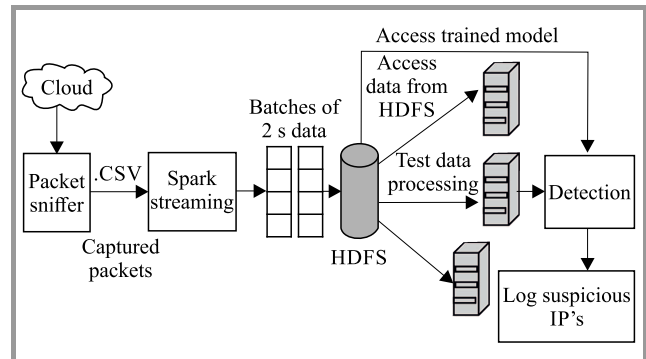


Fig. 3. Proposed DDoS attack detection system using Spark.

communicate with each other. Packet sniffer collects network traffic and identifies information pertaining to each packet. This raw data (packets) is sent to the Spark processor for feature extraction, as shown in Fig. 3.

Algorithm 2 is designed for capturing network traffic in the Neutron node of the OpenStack cloud. Initially, the socket library is imported and the creation of a raw socket follows. It listens on a particular port of the neutron node's IP, while the other socket reaches out to other VMs to form a connection. The listener socket is formed at the server while the client reaches out to the server. Packet sniffer gathers raw binary data from the connection established. The selected network interface is switched to promiscuous

Algorithm 2. Capturing of packets using packet sniffer**Input:** Neutron node IP interface**Output:** CSV file

- 1: Create raw socket
- 2: Bind a raw socket to Eth0 interface of neutron node
- 3: Set the interface to promiscuous mode
- 4: Capture data link layer (Ethernet) header
- 5: Retrieve IP/ICMP/TCP header fields by parsing Ethernet frame
- 6: Extract the features as per Table 1
- 7: Write the extracted features to CSV file

mode. Promiscuous mode is used to direct all network traffic through the chosen network interface. It captures various fields of the chosen protocol headers. The packet sniffer module is designed to collect the captured network data and to extract packet information by verifying its protocol. The packet sniffer module captures incoming packets for a required period of time. All information pertaining to the protocol headers is stored as a CSV file. A DDoS attack is simulated in the cloud environment with a view to sending packets with large amounts of data, and to from multiple connections to simulate connection flooding and to captures the attack. The features captured in this process are shown in Table 1.

Table 1
Dataset features captured

Features	
Source MAC	Acknowledgement number
Destination MAC	SYN bit
Protocol	ACK bit
Service	FIN bit
Source address	Src bytes
Destination address	Time stamp
Sequence number	

3.4. Processing of Captured Data

Real-time data captured in the previous module needs to be processed. The processing of a dataset is performed to obtain the specific features required for testing the model. The distributed computing approach is introduced to accelerate the process. Spark is one of the distributed frameworks that enable processing based on in-memory computing. Initially, the features are recorded once every millisecond. Additionally, the derived features need to be computed for the past two seconds of the connection. The data stream captured by the sniffer module is pre-processed in a batch of 2 s. The algorithm is designed to perform certain actions for feature selection, as not all features that are captured contribute to DDoS detection. Thus, it is important to select the required features. The pre-processor sends this information to the detection engine. Later, the

dataset is processed in a big data environment using the Spark engine.

Algorithm 3. Processing of the captured data**Input:** Generated log file CSV**Output:** Processed test dataset

- 1: Count = number of connections to the same host
- 2: Src_bytes = number of data bytes from source to destination
- 3: Dest_bytes = number of data bytes from destination to source
- 4: Srv_count = number of connections to the same service
- 5: Same_srv_rate = percentage of connections to the same service
- 6: Initialize Spark context and Spark streaming context
- 7: Stream captured data every 2 s
- 8: Read the captured data as dstream
- 9: **For** every RDD from dstream
- 10: Convert RDD to dataframe
- 11: Initialize Count, Src_bytes, Dest_bytes, Srv_count, Same_Srv_rate to zero
- 12: *//Apply map and reduce to get the features*
- 13: **For** each connection i.e. combination of protocol type and service
- 14: Apply map and reduce technique to compute the Count, Src_bytes, Dest_bytes, Srv_count, Same_Srv_rate
- 15: **End for**
- 16: Append all computed features along with protocol type and service to data frame
- 17: Write transformed dataframe to a file
- 18: **End for**

The processing of captured data collected through the packet sniffer is represented in Algorithm 3. This is basically performed for the test dataset collected and the results are used for further detection analysis. Distributed processing is carried out on the Spark framework. The packets sniffed over a specific duration serve as an input for this module. The processing is performed for each 2 s intervals of data in order to analyze network traffic. Initially, Spark context and configuration are created. The packet log data file is read and converted as a dstream object. Finally, each RDD of dstream is converted into a Spark data frame and the selected features are computed using the map reduce function (Protocol_type, Service, Src_bytes, Dest_bytes, Count, Srv_count, Same_srv_rate).

3.5. Classification using ML Algorithms

Machine learning is an effective technique allowing to detect any anomaly-based attacks. Real-time detection of a DDoS attack in a cloud environment requires a trained model. This module builds a trained model using a real-time dataset. The real-time training dataset is generated by recording network activity over the period of 24 hours by simulating an ICMP flooding attack. We compare three

classification algorithms based on accuracy, precision and false-positive rate parameters and choose the best classifier for our model. Such classification algorithms as random forest, decision tree and logistic regression are used for modeling the training dataset. We apply these classification algorithms to both the benchmark dataset and the real-time dataset. Later, we use the Spark MLlib library to implement the selected classification algorithm. As the Spark framework is designed to work in a distributed environment, we use a distributed file system. The selected training model is stored in HDFS for further retrieval. This trained model is accessed by the detection module to detect attacks as shown in Fig. 3.

3.6. Detection Module

The process of identifying anomalous attacks in normal traffic is known as detection. Existing cloud behaviors related to attacks are modeled using the trained model. As shown in Algorithm 4, a real-time test dataset is generated using the packet sniffer, by simulating a DDoS attack. This test dataset is processed using Spark streaming. The captured data is considered as a test dataset and is fed into the trained classifier. If a connection is considered to constitute an attack, then we log the suspicious IPs. The administrator checks the suspicious IP addresses and can observe the traffic anomaly over a period of time and may block the IP addresses. All four modules are implemented on the private cloud testbed.

Algorithm 4. Detection module

Input: Processed test dataset

Output: List of suspicious IP's

- 1: Capture the network packets for training using packet sniffer
 - 2: Compute the features using Spark streaming
 - 3: Read processed test dataset
 - 4: Feed the processed test dataset into trained model to check for intrusion
 - 5: **If** intrusion is detected **Then**
 - 6: Log the suspicious IP's for admin analysis
 - 7: **End if**
-

4. Results and Discussions

In this section, we describe the configuration of the experimental OpenStack-based cloud testbed. We also discuss the results obtained while selecting the machine learning classification algorithm. Finally, we discuss the training and prediction times for the selected ML classifier, execution time for detection of multiple slave nodes, and duration for capturing various packets.

4.1. OpenStack Cloud Environment

OpenStack [20], is a real-time cloud environment used to capture network data and to classify attacks. The private

cloud setup comprises 5 nodes, namely a controller node, a neutron node and three compute nodes. The controller node provides OpenStack services and APIs to carry out the various tasks performed by the network and by the compute nodes. The neutron node helps in networking services, such as DHCP which assigns IPs to virtual machine instances. The compute node provides VM nodes using a hypervisor. Details of the real-time private cloud setup are described in Table 2.

Table 2
OpenStack testbed configuration

Nodes	IP Address	Memory
Controller	192.168.31.2	1 GB
Neutron	192.168.31.3	1 GB
Compute node 1	192.168.31.4	4 GB
Compute node 2	192.168.31.5	4 GB
Compute node 3	192.168.31.6	4 GB

The packet sniffer module was designed for capturing network traffic. The objective of the analysis is to understand the behavior experience during an attack behavior on the private cloud testbed.

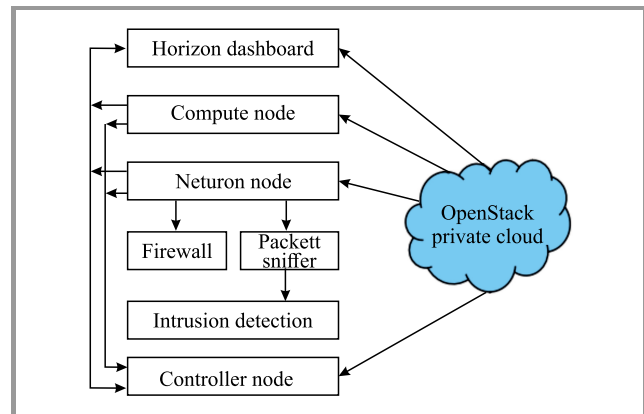


Fig. 4. Integration of packet sniffer and IDS on OpenStack.

Figure 4 shows the integration of the packet sniffer and IDS in an OpenStack cloud. The packet sniffer module is installed on the neutron node of the cloud to capture activities of the cloud. An ICMP flooding attack is performed to simulate attacks on the cloud. The behavior of the cloud is captured. The intrusion detection system designed on the spark cluster is deployed in the cloud to detect any suspicious activity.

4.1.1. Apache Spark Platform

In this experiment, we have used one master node and varying numbers of slave nodes (up to 3). We used Apache Spark [21] for distributed processing. Spark clusters are generated with the use of OpenStack virtual machines, relying on the automated process described earlier. It should

be noted that the configuration in both the slave machines and the master machine are the same. The Spark configuration has been set as follows:

- Spark executor instances = 4,
- Spark executor cores = 10,
- size of Spark executor memory = 20,
- size of Spark driver memory = 9 B.

As the Spark framework is designed to work in a distributed environment, the Hadoop distributed file system (i.e. HDFS) was adopted.

4.2. Selection of Machine Learning Classifier

We use three parameters to select the best machine learning classifier:

Accuracy. Accuracy is the percentage ratio between the number of correctly classified samples from the dataset and the total number of samples of the dataset, and it reflects the discrimination capability of the classifier:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%$$

where TP – true positive, TN – true negative, FP – false positive, FN – false negative.

False Positive Ratio (FPR). FPR indicates the proportion between wrong classified normal samples and all normal samples. FPR is a parameter that indicates the probability of falsely rejecting the null hypothesis for a particular test:

$$FPR = \frac{FP}{FP + TN}$$

Precision. Precision is defined as the number of true positives divided by the number of true positives with the addition of the number of false positives. Precision describes the ability of a classification model to return relevant instances only:

$$Precision = \frac{TP}{FP + TP} \cdot 100$$

Classification is a major factor in identifying an intrusion attack. It is essential to deploy the best data model to obtain accurate classification results. Table 3 presents a com-

parison of performance of 3 ML classifiers, namely random forest, decision tree, and logistic regression. Furthermore, accuracy, false positive rate and precision are defined for two types of datasets. The real time dataset is used as training dataset and cross validation is applied. Accuracy is noted for all 3 algorithms, i.e. random forest, decision tree and logistic regression. For KDD Cup 99 dataset the same pattern of accuracy is observed, but the accuracy level being higher than in the case of the real time dataset. From Fig. 5, one may observe that the random forest model offers better accuracy and precision, which renders it suitable for detection.

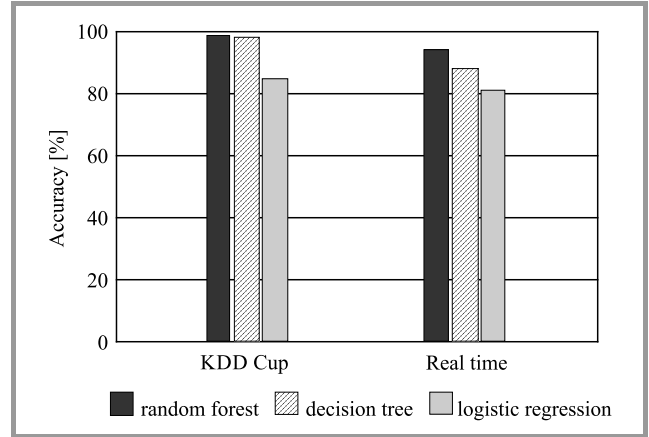


Fig. 5. Accuracy of ML models for two datasets.

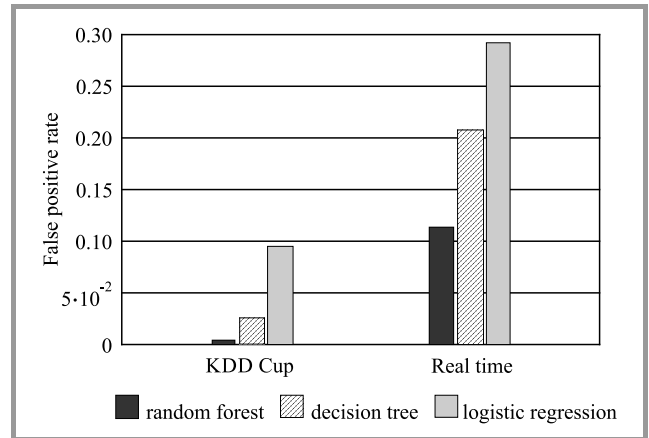


Fig. 6. False positive rate of ML models for two datasets.

The false positive ratio (FPR) is the most important parameter in anomaly detection. As the traffic pattern in the cloud network changes along with the varying intervals and events, traffic bursts are experienced. Thus, a sudden change in traffic volume may result in more FPR. This may trigger a lot of alerts for the administrator. Thus, choosing a classifier with better FPR is important in DDoS attack detection. As shown in Fig. 6, the value of FPR for the KDD Cup 99 dataset is 0.003 for random forest, 0.022 for decision tree and 0.095 for logistic regression algorithms, respectively. The value of FPR for the real time dataset is 0.111 for random forest, 0.208 for decision tree and 0.293

Table 3

Comparison of three classification models

Dataset	ML model	Accuracy	FPR	Precision
KDD Cup	Random forest	99.21%	0.003	99.91%
	Decision tree	98.82%	0.024	99.37%
	Logistic regression	84.97%	0.095	98.84%
Real time	Random forest	94.40%	0.111	89.87%
	Decision tree	88.24%	0.208	78.74%
	Logistic regression	81.43%	0.293	72.42%

for logistic regression. The rate of detection of false positive samples from benchmark and real time datasets is, in random forest, lower than in the case of logistic regression and decision tree algorithms. In summary, the random forest algorithm performs better in terms of accuracy, FPR, and precision. Thus, we choose this classifier for detection of DDoS attacks in a distributed environment.

4.3. Results using Apache Spark

In this section, we discuss the results obtained using the random forest model, with different parameters and scenarios applied.

The process of providing the random forest model with training data to learn is known as training time. The learning algorithm finds patterns in the training data that map the input data attributes to the target, and it outputs an ML model that captures these patterns. The time taken for this process is training time. Figure 7 provides different training times along different spark cluster sizes for the random forest algorithm. MLlib libraries of Spark provide training and prediction times while training the classifier. The results illustrate processing speed increases as cluster size increases.

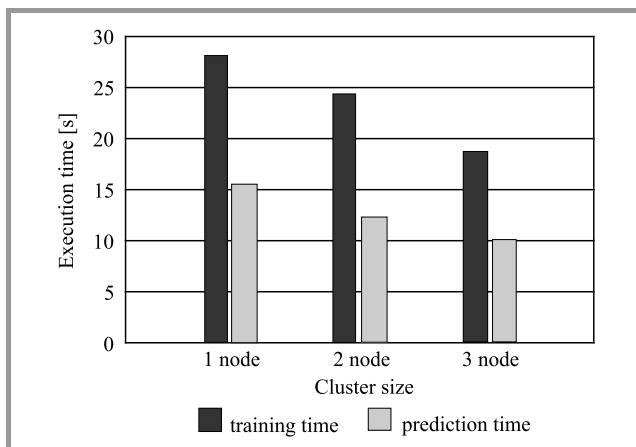


Fig. 7. Training and prediction time vs. different cluster sizes.

Figure 8 shows that as the Spark cluster size increases, the DDoS attack detection time reduces for distributed processing using Spark. The packet sniffer is executed for 4000 s and the CSV file is created and used as an input dataset for the detection of DDoS attacks. The Spark distributed detection algorithm is executed for the random forest classification model and the captured test dataset. The figure shows also a comparison between non-distributed processing and distributed processing using Spark for the scenario in question. The distributed detection process varies with different cluster sizes, such as a single node, 2 node and 3 node master-slave Spark cluster, as shown in Fig. 8.

The detection time is executed in a 2-node Spark cluster for different durations of the packets captured. The packet sniffer is executed for 10, 20 and 30 seconds. 1024 and log CSV files are created. These log files are tested against

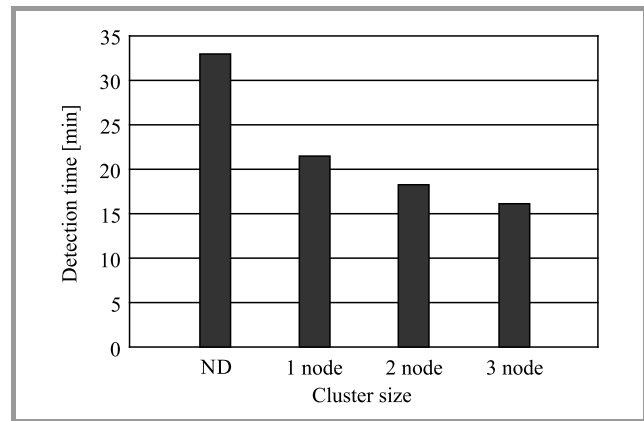


Fig. 8. Detection time vs. different cluster sizes.

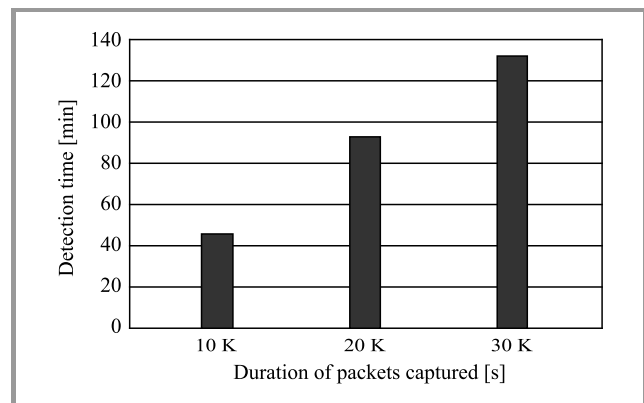


Fig. 9. Detection time vs. duration of captured packets.

the classification model to detect the attacks. The detection time is longer for the larger duration of captured packets. Based on the Fig. 9, we infer that as the capturing time increases, log file size increases. Thus, the execution time increases.

5. Conclusions

This work proposed a system for efficient detection of DDoS attacks in a private cloud environment. Initially, we designed Spark as a service in an OpenStack-based private cloud for on-demand provisioning of Spark clusters. Then, we evaluated three machine learning models using benchmark and real-time datasets and selected the random forest model for the classification of DDoS attacks. Then, we used this algorithm in the distributed environment. By using RDD Spark streaming, we performed the pre-processing, training and testing stages. The results reveal that the time for detecting DDoS attacks is reduced and the detection efficiency is improved significantly thanks to the advantages offered by the Spark framework. OpenStack was used to setup the private cloud with which the intrusion detection system is integrated.

As future work, we plan to use a deep learning model and detect different types of DDoS attacks. Furthermore,

the work may be extended to include parameter tuning of the Spark framework for increased computational efficiency.

References

- [1] S. Dong, K. Abbas, and R. Jain, "A survey on distributed denial of service (DDoS) attacks in SDN and cloud computing environments", *IEEE Access*, vol. 7, pp. 80813–80828, 2019 (DOI: 10.1109/ACCESS.2019.2922196).
- [2] N. Agrawal and S. Tapaswi, "Defense mechanisms against DDoS attacks in a cloud computing environment: state-of-the-art and research challenges", *IEEE Commun. Surv. & Tutor.*, vol. 21, no. 4, pp. 3769–3795, 2019 (DOI: 10.1109/COMST.2019.2934468).
- [3] M. Zekri, S. E. Kafhali, N. Aboutabit, and Y. Saadi, "DDoS attack detection using machine learning techniques in cloud computing environments", in *Proc. 3rd Int. Conf. of Cloud Comput. Technol. and Appl. CloudTech 2017*, Rabat, Morocco, 2017, pp. 1–7 (DOI: 10.1109/CloudTech.2017.8284731).
- [4] C. N. Modi, D. R. Patel, A. Patel, and M. Rajarajan, "Integrating signature apriori based network intrusion detection system (NIDS) in cloud computing", *Procedia Technol.*, vol. 6, pp. 905–912, 2012 (DOI: 10.1016/j.procy.2012.10.110).
- [5] A. S. Syed Navaz, V. Sangeetha, and C. Prabhadevi, "Entropy based anomaly detection system to prevent DDoS attacks in cloud", *Int. J. of Comp. Appl.*, vol. 42, no. 15, pp. 42–47, 2013 [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/1308/1308.6745.pdf>
- [6] R. Karimazad and A. Faraahi, "An anomaly-based method for DDoS attacks detection using RBF neural networks", in *Proc. of Int. Conf. on Netw. and Electron. Engin. IPCSIT 2011*, vol. 11 [Online]. Available: <http://ipcst.com/vol11/9-ICNEE2011-N019.pdf>
- [7] A. Zarrabi and A. Zarrabi, "Internet intrusion detection system service in a cloud", *Int. J. of Comp. Sci. Issues*, vol. 9, iss. 5, no. 2, pp. 308–315, 2012 [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?jsessionid=1F3769647FE79FC6E52BF98BF7C503C?doi=10.1.1.401.7096&rep=rep1&type=pdf>
- [8] H. Gajjar and Z. Malek, "A survey of intrusion detection system (IDS) using OpenStack private cloud", in *Proc. 4th World Conf. on Smart Trends in Syst., Secur. and Sustainab. WorldS4 2020*, London, United Kingdom, 2020, pp. 162–168 (DOI: 10.1109/WorldS450073.2020.9210313).
- [9] J. Choi, C. Choi, B. Ko, D. Choi, and P. Kim, "Detecting Web based DDoS attack using MapReduce operations in cloud computing environment", *J. of Internet Serv. and Inform. Secur.*, vol. 3, no. 3/4, pp. 28–37, 2013 [Online]. Available: <http://isyu.info/jisis/vol3/no34/jisis-2013-vol3-no34-03.pdf>
- [10] M. Mizukoshi and M. Munetomo, "Distributed denial of services attack protection system with genetic algorithms on Hadoop cluster computing framework", in *Proc. of IEEE Congr. on Evolut. Comput. CEC 2015*, Sendai, Japan, 2015, pp. 1575–1580 (DOI: 10.1109/CEC.2015.7257075).
- [11] A. Alsirhani, S. Sampalli, and P. Bodorik, "DDoS attack detection system: utilizing classification algorithms with Apache Spark", in *Proc. 9th IFIP Int. Conf. on New Technol., Mobil. and Secur. NTMS 2018*, Paris, France, 2018, pp. 1–7 (DOI: 10.1109/NTMS.2018.8328686).
- [12] O. Osanaiye *et al.*, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing", *EURASIP J. on Wirel. Commun. and Network.*, vol. 2016, article no. 130, 2016 (DOI: 10.1186/s13638-016-0623-3).
- [13] A. Bhatia, "Faster detection and prediction of DDoS attacks using MapReduce and time series analysis", in *Proc. Int. Conf. on Inform. Network. ICOIN 2018*, Chiang Mai, Thailand, 2018, pp. 556–561 (DOI: 10.1109/ICOIN.2018.8343180).
- [14] S. Hameed and U. Ali, "HADEC: Hadoop-based live DDoS detection framework", *EURASIP J. on Inform. Secur.*, vol. 2018, article no. 11, 2018 (DOI: 10.1186/s13635-018-0081-z).
- [15] A. Sharma, C. Agrawal, A. Singh, and K. Kumar, "Real-time DDoS detection based on entropy using Hadoop framework", in *Computing in Engineering and Technology. Proceedings of ICCET 2019*, B. Iyer, P. S. Deshpande, S. C. Sharma, and U. Shiurkar, Eds. AISC, vol. 1025, pp. 297–305. Springer, 2019 (DOI: 10.1007/978-981-32-9515-5_28).
- [16] C. Wang, T. T. Miu, X. Luo, and J. Wang, "Skyshield: a sketch-based defense system against application layer DDoS attacks", *IEEE Trans. on Inform. Forensics and Secur.*, vol. 13, no. 3, pp. 559–573, 2018 (DOI: 10.1109/TIFS.2017.2758754).
- [17] S. Behal, K. Kumar, and M. Sachdeva, "D-FACE: an anomaly based distributed approach for early detection of DDoS attacks and flash events", *J. of Netw. Comp. Appl.*, vol. 111, pp. 49–63, 2018 (DOI: 10.1016/j.jnca.2018.03.024).
- [18] K. B. Virupakshar, Narayan D. G., and P. S. Hiremath, "Detection of DDoS attacks in software defined networks", in *Proc. of 3rd Int. Conf. on Computat. Sys. and Inform. Technol. for Sustain. Solut. CSITSS 2018*, Bengaluru, India, 2018, pp. 265–270 (DOI: 10.1109/CSITSS.2018.8768551).
- [19] K. B. Virupakshar *et al.*, "Distributed denial of service (DDoS) attacks detection system for OpenStack-based private cloud", *Procedia Comp. Sci.*, vol. 167, pp. 2297–2307, 2020 (DOI: 10.1016/j.procs.2020.03.282).
- [20] OpenStack [Online]. Available: <https://www.OpenStack.org/>
- [21] Apache Spark [Online]. Available: <https://spark.apache.org>



Shweta Gumaste received her B.E. in Computer Science and Engineering from KLE Technological University, Hubballi, India, in 2019. Currently, she is working as associate software engineer in Robert Bosch Engineering and Business Solutions, Bangalore.

E-mail: shwetagumaste@gmail.com
School of Computer Science and Engineering
KLE Technological University
Hubballi, Karnataka, India



Narayan D. G. is currently working as a Professor in the School of Computer Science and Engineering at KLE Technological University, Hubballi, India. He obtained Ph.D. and M.Tech. from Visvesvaraya Technological University, Belgaum in 2017 and 2004 respectively. He obtained B.E. in Computer Science and Engineering from Karnataka University in 2000. He has 20 years of teaching and research experience. His research interest includes wireless mesh networks, cyber security, software defined networks and blockchain. He is a reviewer for many SCIE indexed journals. He has published more than 35 papers in conferences and journals.

 <https://orcid.org/0000-0002-2843-8931>

E-mail: narayan_dg@kletech.ac.in

School of Computer Science and Engineering

KLE Technological University

Hubballi, Karnataka, India



Sumedha Shinde received her B.Sc. and M.Sc. degrees in Mathematics from Karnataka University in 1998 and 2000 respectively. She obtained Ph.D. in Spectral graph theory in 2017 from Visvesvaraya Technological University, Belgaum. Currently, she is working as an Assistant Professor in Department of Mathematics at

KLE Technological University, Hubballi, India. Her research interest includes spectral graph theory, application of graph theory in computer networks and applied statistics. She has 20 years of experience and published more than 10 papers in conferences and journals.

 <https://orcid.org/0000-0002-4154-0025>

E-mail: sumedha@kletech.ac.in

Department of Mathematics

KLE Technological University

Hubballi, Karnataka, India



Amit K holds his M.Sc. degree in Computer Science and currently pursuing Ph.D. in the area of network security. He is working as an Assistant professor in KLE Technological University, Hubballi, Karnataka and is having teaching experience of 12 years. His research interest includes network security, artificial intelligence and cloud computing. He has published papers in the area of IoT and teaching pedagogy in international conferences.

E-mail: amitk@kletech.ac.in

Department of MCA

KLE Technological University

Hubballi, Karnataka, India

A New Parameterized Model for Determining Quality of Online Video Service Using Modern H.265/HEVC and VP9 Codecs

Tadeus Uhl¹ and Christian Hoppe²

¹ Maritime University of Szczecin, Szczecin, Poland

² Nextragen Solutions GmbH, Handewitt, Germany

<https://doi.org/10.26636/jtit.2020.146220>

Abstract—This paper describes a new measurement method (VS model) for determining the quality of online video services relying on modern H.265/HEVC and VP9 codecs. The said method has been developed on the basis of VQuad-HD curves (according to ITU-T J.341). This model does not refer to signal analysis, but protocol analysis instead. The parameters used are: type of video codec, encoding rate, transport technique, packet loss and burst size. The method may be implemented quickly and easily, which is one of the great advantages when using this method to measure QoS.

Keywords—H.265/HEVC, IP network, ITU-T, J.341, native RTP, QoS measurement techniques, video service, VP9.

1. Introduction

In current IP networks, new types of services continue to appear. Many of them are related to transporting video content that constitutes a major portion of today's Internet traffic. According to CISCO reports [1]–[2], in 2017, video material accounted for 75% of global Internet traffic and for nearly 60% of mobile traffic. The volume of video traffic is growing rapidly, especially in mobile communications, and is estimated to reach nearly 80% of the world's mobile traffic by 2022. It is expected that mobile video will increase nine-fold between 2017 and 2022. Therefore, it makes sense to take a closer look at video services.

Video services are currently rendered in two forms. The first category includes real-time services, e.g. television over IP (IPTV), video telephony over IP (VToIP). The other form encompasses services in which time does not play a principal role. These include video streaming services, e.g. video on demand (VoD). To guarantee a good-quality video service, effective transport techniques are needed. In real-time services, MPEG-2 transport streams based on ISO/IEC 13818-1 [3] and native RTP based on IETF RFC 3640 [4] are currently the most widely used techniques. In video streaming (non-real-time service), native RTP and MPEG-DASH based on ISO/IEC 23009-1:2012 [5] are currently the most widely used techniques. It is worth noting that the native RTP

technique is used in both forms of video service. This is a sufficient reason to devote a complete paper to this particular transport technique that will serve as a basis of all investigations described herein.

In order to ensure high video service image quality on the one hand and to reduce the coding rate (and thus the required bandwidth in the network) on the other, improvements in image coding have been promoted in recent years. A new ITU-T standard H.265/HEVC (high-efficiency video coding) was published in 2013 [6]. A few years later, a new proprietary solution (Google), the VP9 Codec, was launched [7]. Image resolution is also increasing continuously, with full HD succeeding the HDTV format. New formats, 2K and 4K, are gaining in popularity as well. The impact that the developments mentioned here exert on the quality of video services must be measured in practice.

Quality of service (QoS) and quality of experience (QoE) should be continuously monitored in modern networks, preferably automatically. Dedicated measuring systems and methods are obviously indispensable for that purpose. When it comes to video communications, several methods are already available for measuring QoS and QoE. Most of them have been standardized, e.g. ITU-T Rec. J.144 [8], ITU-T J.147 [9], ITU-T J.247 [10], ITU-T J.341 [11], ITU-T P.1204.4 [12]. All these standards have evolved from signal-based QoE measuring methods with “full reference” (intrusive measurement necessary). That is what makes these measuring techniques so complicated, time-consuming, and expensive due to the licensing costs involved.

The second group of QoS measurement methods is based on parameterized models. They are simple, faster to execute and generally more accessible (without licenses). However, they do not provide as accurate results as the QoE measurement methods mentioned above. The most important parameterized QoS models for video services include: mean squared error (MSE) [13], new quality index (NQI) [14] and structural similarity index (SSIM) [15]. A comparison of performance of the above-mentioned QoS techniques was presented in papers [16], [17] and else-

where. Specialized QoS models exist as well, e.g. for VoIP [18] or for IPTV services [19]. The last two QoS models were developed for the “old” video codecs (e.g. H.264/AVC) and for lower picture resolutions (up to SDTV). There are currently no QoS models that do justice to the latest developments in the field of video codecs (such as H.265/HEVC, VP9) and picture formats (such as HDTV, Full HD). The work described in this paper aims to remedy that deficiency.

After the introduction presented above, a new video service (VS) model will be formulated and illustrated in the context of actual video codecs and picture formats. This model is then analyzed in various applications and its practicability is put to the test in a comparison study. The results obtained based on the analyses are presented graphically and are interpreted. The paper is concluded with a summary and an outlook on future work.

2. New Parameterized VS Model

It is generally known that packet losses are some of the worst impairment parameters in the IP environment. This is true in the case of both audio and video transmission. Other parameters that greatly affect QoS values include codec type, encoding rate and burst size. The size of jitter buffers in the terminal equipment has a significant influence on QoS as well. The effects of jitter and out-of-order packet delivery are converted into losses, bearing in mind that these errors may be smoothed out with the aid of a jitter buffer. Two types of packet losses and burst size, i.e. burst ratio, being the ratio of the measured burst size to the burst size expected in a burst-free environment, can be determined by means of practical measurements during an RTP session. Protocol analysis may also determine the video codec, the encoding rate and the type of transport technique used. This is one of the prerequisites enabling the parameterized QoS model to be used in practice.

To investigate the effects of packet loss, burst size and encoding rate on the quality of video services, several extensive numerical calculations, which took several weeks, were carried out within the scope of this paper. The QoSCalc (IPTV) numerical tool [20] was used. The tool relies on the VQuad-HD algorithm [21] to determine the QoS (“full reference” technique). Two modern codecs were selected for the investigation scenarios, i.e. H.265/HEVC and VP9 (video sequence in full HD resolution of 1920×1080). The native RTP method was used here as the transport technique. This technique requires the UDP protocol within the transport layer. This combination of protocols is most frequently used for real-time web-based video services (e.g. VoIP, IPTV).

Figures 1 and 2 show examples of numerical results for the H.265/HEVC codec, i.e. the packet loss and burst size values assumed for the investigation. They show that all QoE curves progress exponentially. It is evident that as packet losses increase, the curves fall rapidly and then run asymptotically at a value between 1 and 1.5 MOS. It is also

evident that burst size has an equally considerable impact on QoE values, with a burst size of “1” producing the worst quality of service in all three cases and increases in burst size leading to improvements in service quality. This is in keeping with the psycho-visual model of the human visual perception system, which asserts that viewers will accept a few isolated major disruptions in reception far more readily than numerous, regularly occurring disturbances of minor character. The QoE curves shown in Figs. 1 and 2, supplemented by curves for further encoding rates, form the basis for creating new parameterized QoS models.

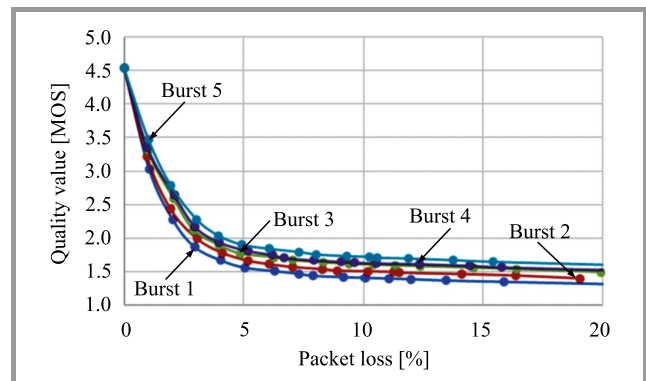


Fig. 1. VQuad-HD values as a function of packet loss and burst size at an encoding rate of 3875 kbps for the H.265/HEVC codec and native RTP.

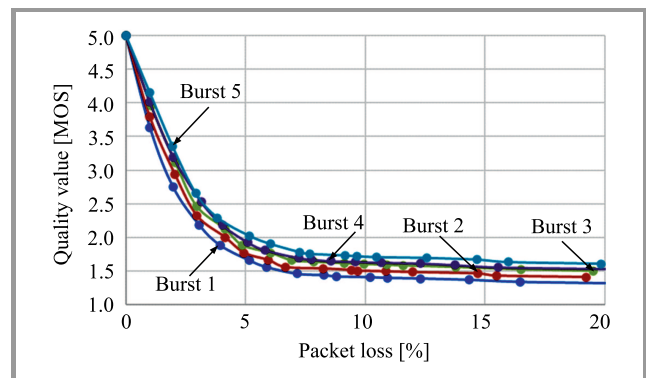


Fig. 2. VQuad-HD values as a function of packet loss and burst size at an encoding rate of 6750 kbps for the H.265/HEVC codec and native RTP.

To develop a new, parameterized QoS model for video services using the modern H.265/HEVC and VP9 codecs, the IPTV model from paper [19] was used. The new parameterized model is referred to as a video service (VS) model in the context of this paper. In addition, the procedure for creating the QoS model from work [19] is cited below and has been used.

Step 1. Using a suitable tool, determine the SwissQual VQuad-HD [21] curves as a function of packet losses, burst size and encoding rate for various codecs and types of transport streams. These curves serve as a basis for further calculations.

Step 2. Approximate each of the SwissQual VQuad-HD curves as a function of packet losses for burst sizes of “1” to “5” and selected encoding rates using:

$$VSfactor = P \cdot e^{\frac{a \cdot packetloss}{burstsize}} + Q \cdot e^{\frac{b \cdot packetloss}{burstsize}}. \quad (1)$$

Constants a and b are selected so that they display values equal to or lower than zero, with significantly smaller values being chosen for b . The result of this is that the two summands are responsible for the steepness of the curves when packet losses are few (second summand) or many (first summand). All constants (P , Q , a , and b) are now calculated iteratively as best possible values for each encoding rate.

Step 3. Determine the formulas the constants P , Q , a , and b . Once determined, the constants are recorded as functions of the encoding rate. The corresponding formulas can then be calculated by means of polynomial approximation. The degree of the polynomial is determined by the complexity of the curve. Equations (2) to (5) show the simple relationship:

$$P = w_n \cdot Bitrate^n + w_{n-1} \cdot Bitrate^{n-1} + \dots + w_1 \cdot Bitrate^1 + w_0, \quad (2)$$

$$Q = z_n \cdot Bitrate^n + z_{n-1} \cdot Bitrate^{n-1} + \dots + z_1 \cdot Bitrate^1 + z_0, \quad (3)$$

$$a = x_n \cdot Bitrate^n + x_{n-1} \cdot Bitrate^{n-1} + \dots + x_1 \cdot Bitrate^1 + x_0, \quad (4)$$

$$b = y_n \cdot Bitrate^n + y_{n-1} \cdot Bitrate^{n-1} + \dots + y_1 \cdot Bitrate^1 + y_0. \quad (5)$$

In the study, for reasons of compatibility with paper [19], two terms were used in Eq. (1). The optimization steps carried out as part of the work described in this paper, e.g. three or four terms in Eq. (1) have not contributed to any improvement in the mode of operation of the model.

Figures 3 to 5 show the parameters as a function of the encoding rate and the results obtained from the approximations made in Eqs. (2) to (5) in case of the H.265/HEVC codec.

Therefore, the actual equations for the H.265/HEVC codec and the native RTP are:

$$P = 3.8 \cdot 10^{-14} \cdot bitrate^3 + 2.79 \cdot 10^{-10} \cdot bitrate^2 - 2.37 \cdot 10^{-5} \cdot bitrate + 1.82, \quad (6)$$

$$Q = 1.71 \cdot 10^{-12} \cdot bitrate^3 - 5.81 \cdot 10^{-8} \cdot bitrate^2 + 6.34 \cdot 10^{-4} \cdot bitrate + 1.14, \quad (7)$$

$$a = -6,36 \cdot 10^{-11} \cdot bitrate^2 + 1.42 \cdot 10^{-6} \cdot bitrate - 2.53 \cdot 10^{-2}, \quad (8)$$

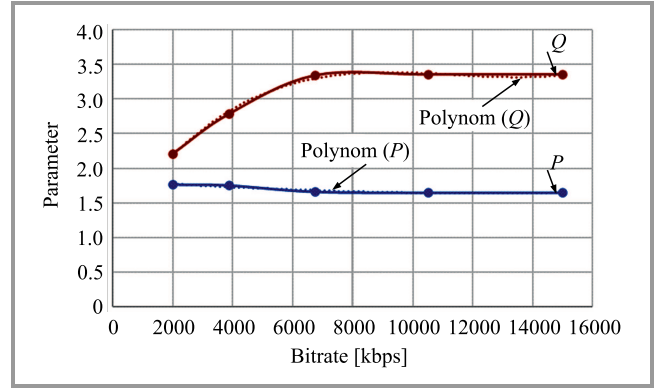


Fig. 3. Approximation of parameters P and Q as functions of the encoding rate for the H.265/HEVC codec and the native RTP.

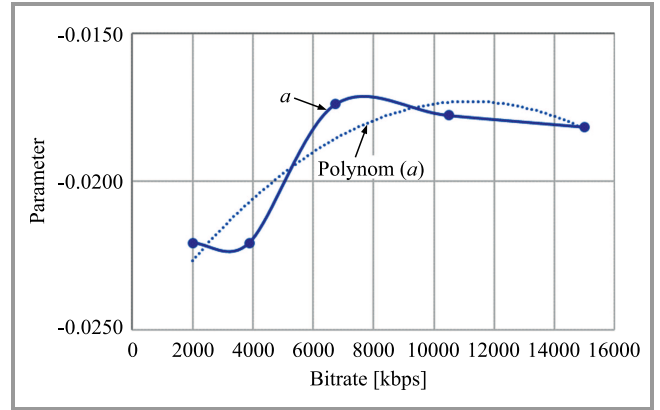


Fig. 4. Approximation of the parameter a as a function of the encoding rate for the H.265/HEVC codec and the native RTP.

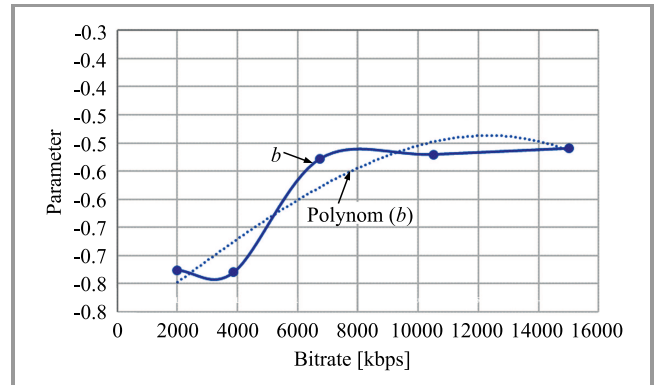


Fig. 5. Approximation of the parameter b as a function of the encoding rate for the H.265/HEVC codec and the native RTP.

$$b = -1.04 \cdot 10^{-13} \cdot bitrate^3 + 3.34 \cdot 10^{-10} \cdot bitrate^2 + 3.93 \cdot 10^{-5} \cdot bitrate - 0.827. \quad (9)$$

For the clarity of presentations, the graphics with QoE values and with parameters P , Q , a , and b are not presented here in the case of the VP9 codec. The calculation performed resulted in the following equations for the VP9 codec and the native RTP:

$$P = -2.89 \cdot 10^{-13} \cdot \text{bitrate}^3 + 4.79 \cdot 10^{-9} \cdot \text{bitrate}^2 - 1.41 \cdot 10^{-5} \cdot \text{bitrate} + 1.66, \quad (10)$$

$$Q = 1.90 \cdot 10^{-12} \cdot \text{bitrate}^3 - 5.99 \cdot 10^{-8} \cdot \text{bitrate}^2 + 6.24 \cdot 10^{-4} \cdot \text{bitrate} + 1.18, \quad (11)$$

$$a = 2.45 \cdot 10^{-14} \cdot \text{bitrate}^3 - 5.28 \cdot 10^{-10} \cdot \text{bitrate}^2 + 2.67 \cdot 10^{-6} \cdot \text{bitrate} - 2.14 \cdot 10^{-2}, \quad (12)$$

$$b = -6.04 \cdot 10^{-14} \cdot \text{bitrate}^3 + 1.30 \cdot 10^{-10} \cdot \text{bitrate}^2 + 3.69 \cdot 10^{-5} \cdot \text{bitrate} - 0.854. \quad (13)$$

The combination of Eqs. (1) and (6) to (9) describes the new, parameterized VS model for the H.265/HEVC codec. In contrast, the combination of Eqs. (1) and (10) to (13) forms the new, parameterized VS model for the VP9 codec. The two QoS models developed here complement the existing range of parameterized QoS models from papers [18] and [19] well, which is of great benefit to any practical utilization.

It should be noted that the new QoS model was determined for encoding bitrates up to 15,000 kbps. In practice, the new model is effective only in this range. The encoding rates chosen in this study correspond to the values that are used in the Internet (e.g. by the Netflix [22]) for video streaming. It should also be noted that in the new model, the following units are to be used for the variables used there: packet losses in the unit [%] and encoding bitrate in the unit [kbps].

Below is an example of the application of the newly created QoS model. The following assumptions are made: H.265/HEVC codec, encoding rate equal to 5000 kbps. The parameters determined according to Eqs. (6)–(9) are: $P = 1.71$, $Q = 3.07$, $a = -0.0197$, $b = -0.635$. The QoS values determined for packet losses of 0 to 5% are shown in Table 1.

Table 1

QoS values as a function of packet losses in the assumed example

Packet loss [%]	0	1	2	3	4	5
QoS values [MOS]	4.78	3.3	2.5	2.07	1.82	1.68

3. Comparison Study

The QoSCalc (IPTV) software tool from paper [20] was used in the following analyses as well. Figure 6 shows the block diagram of the numerical tool used in this paper. This tool operates in the following manner:

- a reference video file is loaded,
- the video is encoded in accordance with the selected codec by FFmpeg [23],

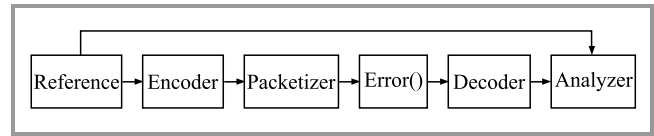


Fig. 6. Schematic representation of quality measurement based on QoSCalc(IPTV).

- the coded data is encapsulated according to the selected transport protocol (here native RTP [24]) by FFmpeg,
- the block “Error” represents the generation of a selected level of network impairments,
- the packed video is decoded to the same format as the reference (raw video, same resolution and encoding rate) by FFmpeg,
- finally, the decoded data and the video reference file loaded at the start are passed on to the evaluation algorithms (here VQuad-HD and VS model).

The “Error” block has been designed for non-deterministically distributed packet loss (binominal distribution with probability P) and non-deterministically distributed burst size (exponential distribution) with a selectable mean value.

The following parameters were assumed for the numerical comparison study:

- encoding: H.265/HEVC (medium); VP9 (FFmpeg default),
- packaging: native RTP with NAL size equal to 1200 bytes,
- full HD video sequence with image refresh frequency of 25 images/s [25],
- encoding rates: 3875, 6750 and 10,500 kbps,
- packet loss (binominal distributed): 0–12 (in increments of 1), 14, 16, 20%,
- burst size: 1, 2, 3 (practical values),
- SwissQual VQuad-HD and VS model as the QoE/QoS measuring techniques,
- 31 measurements for each determined performance value. This number of tests ensures a confidence interval of less than 10% of the estimated average (with a 5% error probability).

The results of the comparison study are presented in Figs. 7 to 12. Figures 7 to 10 refer to the H.265/HEVC codec. Here, two extremely different coding rates (3875 and 10,500 kbps) and burst sizes (1 and 3) were deliberately chosen to demonstrate the effectiveness of the new model in the entire range of its applicability.

Figures 7 to 10 show that QoS deteriorates exponentially as packet losses increase. This is the case for both QoE/QoS

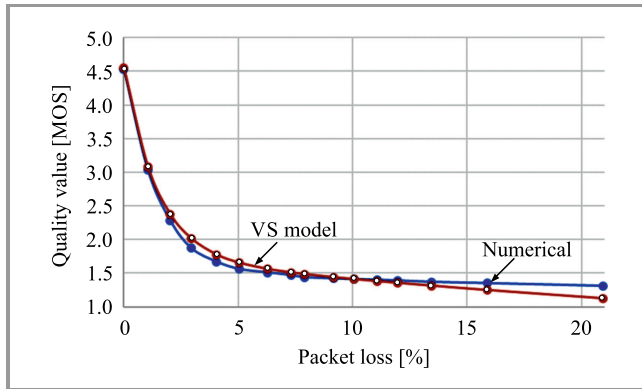


Fig. 7. QoS values as functions of packet losses gained from different measuring methods for the H.265/HEVC codec and the native RTP, a burst size of “1” and an encoding rate of 3875 kbps.

measuring techniques used. Furthermore, the curves fall less steeply as burst size increases. The reason for this is that synchronization of I/P/B images fails more frequently when small groups of packet losses occur regularly, rather than when large groups of packet losses occur infrequently. Moreover, the more numerous the breakdowns in synchronization are, the more frequently the im-

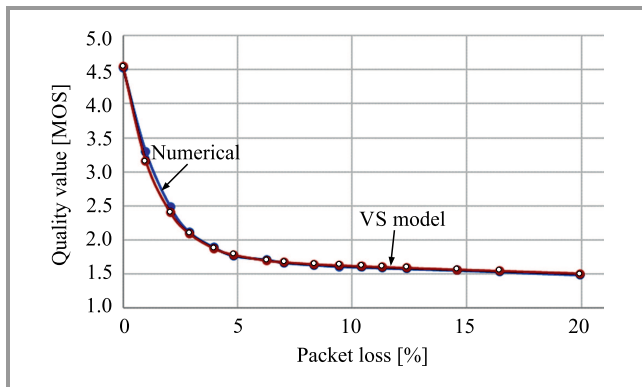


Fig. 8. QoS values as functions of packet losses gained from different measuring methods for the H.265/HEVC codec and the native RTP, a burst size of “3” and an encoding rate of 3875 kbps.

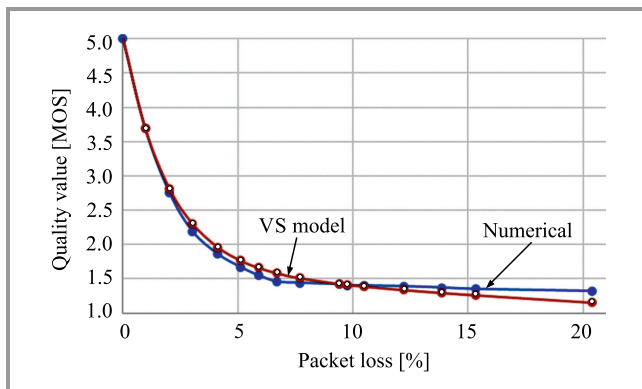


Fig. 9. QoS values as functions of packet losses gained from different measuring methods for the H.265/HEVC codec and the native RTP, a burst size of “1” and an encoding rate of 10,500 kbps.

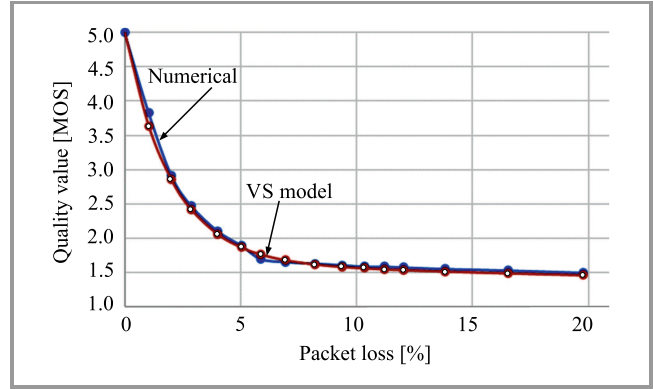


Fig. 10. QoS values as functions of packet losses gained from different measuring methods for the H.265/HEVC codec and the native RTP, a burst size of “3” and an encoding rate of 10,500 kbps.

ages will freeze. This will naturally be reflected in a drop in QoE/QoS values.

Figures 7 to 10 also show that the curves of the VQuad-HD and the VS model progress very closely to each other. Only when burst size is “1” and large packet losses (16–20%) occur, relatively large deviations in the QoS values may be observed. However, they can be accepted from a practical point of view. In practice, such large packet losses indicate a network fault, which in turn leads to the connection being broken. In other words: the numerical comparison study has delivered strong arguments for using the VS model in everyday practice.

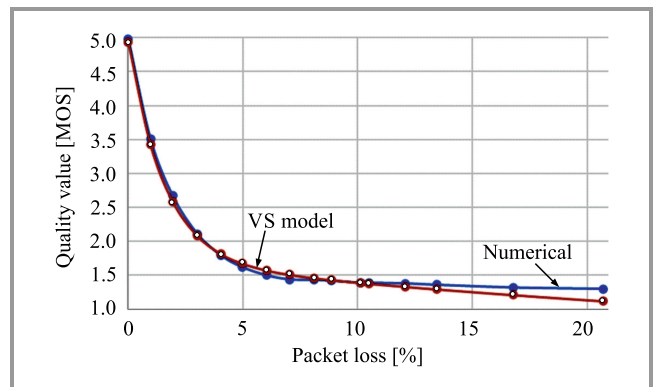


Fig. 11. QoS values as functions of packet losses gained from different measuring methods for the VP9 codec and the native RTP, a burst size of “1” and an encoding rate of 6750 kbps.

Figures 11 and 12 show the results of the investigation from this comparative study using the VP9 codec in two examples. In order to show the effectiveness of the new QoS model in comparison with the investigations carried out previously within the scope of this study (H.265/HEVC codec), the burst sizes of 1 and 2, and the bit rate of 6750 kbps were deliberately chosen. A considerable agreement between the numerical values and the values from the VS model has been observed. This once again confirms good practical functionality of the VS model.

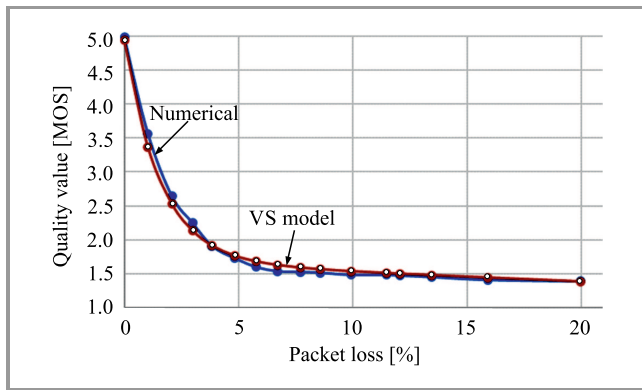


Fig. 12. QoS values as functions of packet losses gained from different measuring methods for the VP9 codec and the native RTP, a burst size of “2” and an encoding rate of 6750 kbps.

4. Summary and Outlook

In the course of the work described in this paper a new parameterized QoS measuring model for two new video codecs (H.265/HEVC and VP9) was developed to determine the quality of real-time video service over IP network, and its functionality was put to the test in a comparison study. The new VS model is based on the VQuad-HD curves. This is of special practical importance, since the VQuad-HD algorithm (according to ITU-T J.341) is the most objective QoE measuring method for video. The comparison study has proved, without a shadow of doubt, practical suitability of the new QoS model. The new, inexpensive, quick and easy-to-use VS model (offline) is far more than simply an adequate alternative to the laboriously slow (online) VQuad-HD method with its expensive licenses.

It would be extremely useful to verify the results presented in this paper in a subjective study (which is unfortunately, extremely time-consuming and resource-intensive). This is planned in the future.

Development in the field of video coding continues. First announcements of the new H.266/VVC standard may already be found [26]. Image resolution is also being increased on a continuous basis (2K and 4K). This should also be considered in further studies and developments. This would enable the currently available range of parameterized QoS models for live video streaming to be expanded and used in practice. In the new developments of such QoS models, one could very well apply the procedure presented in this paper. Further work is planned in this direction, too.

References

- [1] “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022”, Cisco, 2019 [Online]. Available: <https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf>
- [2] “Cisco Predicts More IP Traffic in the Next Five Years Than in the History of the Internet”, Cisco, 2018 [Online]. Available: <https://newsroom.cisco.com/press-release-content?articleId=1955935>
- [3] “Information technology—Generic coding of moving pictures and associated audio information: Systems”, *ISO/IEC 13818-1*, 2000 [Online]. Available: <https://eccc.colorado.edu/~ecen5653/ecen5653/papers/iso13818-1.pdf>
- [4] “RTP Payload Format for Transport of MPEG-4 Elementary Streams”, *IETF RFC 3640*, 2003 [Online]. Available: <http://www.faqs.org/rfcs/rfc3640.html>
- [5] “Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats”, *ISO/IEC 23009-1:2012*, 2012 [Online]. Available: <https://www.iso.org/standard/57623.html>
- [6] “H.265: High efficiency video coding”, *ITU-T* [Online]. Available: <https://www.itu.int/rec/T-REC-H.265>
- [7] “VP9 Bitstream & Decoding Process Specification”, Google [Online]. Available: <https://storage.googleapis.com/downloads.webmproject.org/docs/vp9/vp9-bitstream-specification-v0.6-20160331-draft.pdf>
- [8] “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference”, *ITU-T* [Online]. Available: <https://www.itu.int/rec/T-REC-J.144-200403-I/en/>
- [9] “Objective picture quality measurement method by use of in-service test signals”, *ITU-T* [Online]. Available: <https://www.itu.int/rec/T-REC-J.147-200207-I/en/>
- [10] “Objective perceptual multimedia video quality measurement in the presence of a full reference”, *ITU-T* [Online]. Available: <https://www.itu.int/rec/T-REC-J.247-200808-I/>
- [11] “Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference”, *ITU-T* [Online]. Available: <https://www.itu.int/rec/T-REC-J.341-201603-I/en/>
- [12] “Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information”, *ITU-T* [Online]. Available: <https://www.itu.int/rec/T-REC-P.1204.4-202001-I/en>
- [13] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures”, in *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, 2009 (DOI: 10.1109/MSP.2008.930649).
- [14] Z. Wang and A. C. Bovik, “A universal image quality index”, in *IEEE Signal Process. Let.*, vol. 9, no. 3, pp. 81–84, 2002 (DOI: 10.1109/97.995823).
- [15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity”, *IEEE Transac. on Image Process.*, vol. 13, no. 4, 2004 (DOI: 10.1109/TIP.2003.819861).
- [16] U. Sara, M. Akter, and M. Uddin, “Image quality assessment through FSIM, SSIM, MSE and PSNR – A comparative study”, *J. of Comp. and Commun.*, vol. 7, no. 3, 2019 (DOI: 10.4236/jcc.2019.73002).
- [17] J. Klink and T. Uhl, “Video quality assessment: some remarks on the selected objective metrics”, in *Proc. of the 28th Int. Conf. on Software, Telecommun. and Comp. Networks, IEEE Transaction on Image Process.*, vol. 13, no. 4, pp. 600–612, Split, Croatia, 2004 (DOI: 10.1109/TIP.2003.819861).
- [18] T. Uhl and S. Paulsen, “The new, parameterized VT Model for determining quality in the video telephone service”, *Bull. of the Polish Acad. of Sci.: Techn. Sci.*, vol. 61, no. 3, pp. 431–437, 2014 (DOI: 10.2478/bpasts-2014-0045).
- [19] T. Uhl and H. Juergensen, “The new, parameterized IPTV Model for determining the quality in the IPTV service”, *Bull. of the Polish Acad. of Sci.: Techn. Sci.*, vol. 63, no. 2, pp. 495–500, 2015 (DOI: 10.1515/bpasts-2015-0057).
- [20] T. Uhl and H. Juergensen, “New tool for examining QoS in the IPTV service”, in *Proc. of the World Telecommun. Congress WTC’14*, Berlin, Germany, 2014 (DOI: 10.13140/RG.2.1.3087.5608).
- [21] “SwissQual VQuad-HD Selected as New ITU HD Video Testing Standard J.341” [Online]. Available: <https://www.realwire.com/releases/SwissQual-VQuad-HD-Selected-as-New-ITU-HD-Video-Testing-Standard-J341>
- [22] “Internet Connection Speed Recommendations”, Netflix [Online]. Available: <https://help.netflix.com/en/node/306>

- [23] "FFmpeg project software" [Online]. Available: <http://ffmpeg.org/download.html>
- [24] S. Wenger, M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, "RTP payload format for H.264 Video, RFC 3984" [Online]. Available: <https://datatracker.ietf.org/doc/rfc3984/>
- [25] Opticom Company GmbH, "PEVQ" [Online]. Available: <http://www.opticom.de>
- [26] "H.266 (ex H.VVC) "Versatile video coding" (New)", *ITU-T* [Online]. Available: <https://www.itu.int/md/T17-SG16-200622-TD-PLN-0445/>



Christian Hoppe received his B.Eng. in Communications Technology from the Flensburg University of Applied Sciences, Germany, in 2010 and his M.Sc. in Information Technology from the Kiel University of Applied Science, Germany, in 2016. Today, he works as a Software Engineer for Nextragen Solutions GmbH

in Handwitt, Germany. He focuses mainly on the following areas: quality assurance for triple play services and medical imaging solutions.

E-mail: christian.hoppe@nextragen-solutions.de
Nextragen Solutions GmbH Flensburg
Heideland 1
D-24976 Handewitt, Germany



Tadeus Uhl received his M.Sc. in Telecommunications from the Academy of Technology and Agriculture in Bydgoszcz, Poland in 1975, his Ph.D. from the Gdańsk University of Technology, Poland in 1982 and his D.Sc. from Dortmund University, Germany in 1990. Since 1992, he has been working as a Professor at the Institute of Communications Technology, Flensburg University of Applied Sciences, Germany and, in addition, since 2013, as a Professor at the Institute of Transport Engineering and Economics, Maritime University of Szczecin, Poland.

His activities focus on the following areas: traffic engineering, performance analysis of communications systems, measurement and evaluation of communications protocols, QoS and QoE by triple play services, Ethernet and IP technology. He is an author or co-author of five books and 130 papers dealing with LAN, WAN and NGN.

His activities focus on the following areas: traffic engineering, performance analysis of communications systems, measurement and evaluation of communications protocols, QoS and QoE by triple play services, Ethernet and IP technology. He is an author or co-author of five books and 130 papers dealing with LAN, WAN and NGN.

 <https://orcid.org/0000-0001-6849-9168>

E-mail: t.uhl@am.szczecin.pl
Maritime University of Szczecin
Henryka Pobożnego 11
70-507 Szczecin, Poland

An Online Stream Monitoring Algorithm for Fraud Detection in the Transport of Goods

Paweł M. Białoń

National Institute of Telecommunications, Warsaw, Poland

<https://doi.org/10.26636/jtit.2020.146620>

Abstract— The process of monitoring vehicles used in road transports plays an important role in detecting fraud committed by drivers. Algorithm designers face a number of challenges, including large number of vehicles monitored, demands related to online calculations, and ability to easily explain fraud alarms triggered to supervisors who make final decisions about actions to be taken. In this paper, we propose rather general, lightweight stream, online heuristics. The vehicle's position is periodically controlled by a GNSS device. The algorithm detects potential illegal activities along the route between the origin and the destination. Anomalies in the vehicle's trajectory are detected, based on a multi-resolution analysis of the economy of routes. The economy metric is easily understood and verifiable by controllers. The solution is also capable of identifying clearly suspicious trajectories that popular geofencing approaches would overlook. The scale on which the solution may be adopted is obtained thanks to the stream – like nature of the algorithm: essentially, the resources used do not increase along with the size of the input stream (the number of GNSS frames generated for the vehicle). An experiment illustrating the algorithm's viability is presented as well.

Keywords—*fraud detection, stream processing, transport monitoring.*

1. Introduction and Related Work

The process of monitoring road transport operations [1] is usually understood as using a global navigation satellite system (GNSS) – in more cases GPS – to control the position of vehicles. A vehicle is equipped with an on-board unit (OBU) which identifies its GPS position and sends it, periodically, over the Internet, to a host computer system. An operator checks the position of the vehicle in order to control compliance with predefined routes and time schedules, to ensure the safety of the truck and its crew, and to identify any irregularities, such as extra pickups or drops.

Let us focus on the last activity, i.e. detection of fraud. Usually, commercial vehicles travel between predefined origin and destination locations, transporting goods between these points. A similar situation is dealt with in the case of taxi

operations, where the passenger needs to be transported from a particular starting point to a specific destination. In all such situations, there is a risk that the driver may commit fraud by transporting some additional goods or passengers, on their own initiative, along with the legal load. This may require the driver to visit additional locations along his/her route, to illegally drop or pickup additional goods. Drivers may also take care of their private business while working. The methods used to detect such cases of frauds are based on various concepts.

Such methods may be online- or offline-based. In an online analysis, a suspicious behavior of the vehicle is detected by the system quickly after or even before the forbidden behavior takes place. For example, the system detects that a truck is driving in the wrong direction, or that it takes a lengthy break between the loading/dropping locations. An efficient online detection system may be used for triggering an alarm informing monitoring center staff about such events. They, in turn, may contact the driver or send other employees to check on the suspicious vehicle. An offline analysis consists in analyzing journeys after these have been completed (one-by-one or as a set), in order to detect potential fraud.

The detection methods may be used on different scales. The number of vehicles monitored may grow rapidly in large companies. Monitoring systems may also be operated by governments, countrywide, and may cover considerable number of vehicles. For example, road transports of selected goods have been monitored in Poland since 2017, pursuant to the “Act on the goods transport monitoring system” dated 9 March 2017. The main goal of the solution is to control the movement of goods subject to the imposition of excise duty predominantly fuels. The key objective is to prevent unauthorized or unofficial loading or unloading operations that would result in avoiding the payment of taxes. Pursuant to the aforementioned Act, any operations involving transportation of specific goods must be preceded by a suitable notification submitted via a government-operated online system, known as SENT [2]. The notification contains: a CN (or the Polish PKU) classification code of the goods concerned, their quantity, address of loading

and unloading locations, details of contractors involved, date on which the transport operation begins, vehicle details (mainly the license plate number) and owner's details. As of the end of 2018, vehicles transporting goods monitored with the use of the SENT system must also be monitored by GNSS. Therefore, a vehicle must be equipped with a mobile phone with a suitable tracking app. Alternatively, an on-board unit (OBU) already installed in the vehicle may be used in connection with the host system to collect information required by the government. A single frame of tracking information contains primarily GNSS coordinates of the truck, GNSS time corresponding to the specific position, and identification data (of the OBU). There are two sources of information: SENT declarations and tracking information. Several authorities are allowed to control the vehicles in order to verify, whether the actual status of the vehicle matches the declarations submitted.

From the point of view of efficiency of the monitoring process, it is important how the vehicle's position is analyzed within the company. It seems that manual (visual) supervision seems to be the most common approach. The position of a vehicle is displayed on a map, along with its speed and trajectory. This information is analyzed by the operator. Interestingly, according to practitioners, many systems of this type are based on manual analysis. OBUs obtain GPS information periodically (in predefined intervals, upon a predefined distance has been covered, or of the speed has changed by a specific value). It needs to be borne in mind, however, that it is not always possible to send GPS data (to the host) online. The Internet connection may be temporarily unavailable due to switching between different base stations, the vehicle may be present in coverage holes, or the OBU may be switched off. Therefore, OBUs usually buffer GPS frames and retransmit them after the Internet connection has been regained. Interestingly, it is often the case that the current position is transmitted first, followed by the delayed frames. The current position of the vehicle is more important for the operator than its historical location. Therefore, OBUs do not preserve the order of GPS frames while sending. This indicates (though, clearly, does not determine with certainty) that the analysis performed is either manual or, even if automatic, is based on some simple logic that does not require the true order of GNSS frames. A clear drawback of the manual approach consists in its dependence on the mental and physical condition of humans performing the analysis, who are capable of supervising a small number of vehicles only. In order to cope with this, new systems are usually equipped with automated solutions that are based on some simple logic. The logic uses additional information from vehicle sensors (see [3]). Load sensing capabilities are particularly useful, since they allow to easily detect anomalies in the weight of the commodity, caused by illegal drops or pickups made during the trip. Fraud evidence provided by such sensors is of the solid variety, since it is difficult to explain unplanned changes in load weight. On the other hand, equipping vehicles with specialized sensors is not al-

ways possible due to cost, compatibility and scale-related considerations.

Geofencing [4]–[6] is a good example of the automatic analysis-based approach. The area in which the vehicle may be present is defined, and each violation of the borderline triggers an alarm that calls for specific action to be taken. The permitted region is often defined as a corridor along a predefined route. The logic of geofencing is simple and an alarm may be generated based on analyzing a single frame, independently of the previous frames. The geofencing algorithm does not require a lot of memory. As explained in subsequent sections, one of the main difficulties associated with geofencing consists in the need to deal with a potentially large number of routes between two points within the road network. This problem is faced unless we force the vehicle to use a particular route (which is not always viable, e.g. when transport monitoring systems are operated by governments). Some route variants can be complicated and winding, which makes them clearly uneconomical and fraud-suspicious, but still permitted by geofencing. Moreover, drivers may be required to make small detours to rest areas, service stations or by accident, when they get lost along the way or make a mistake at an intersection. Geofencing might be too sensitive for such scenarios.

Not surprisingly, various machine learning techniques have been deployed to detect anomalies and fraud in transport. Various methods are used here, including statistical approach [7], statistical methods combined with the Dempster-Shafer evidence theory [8], separation through support vector machines [9], just to name a few works. Such approaches usually compare the current route to what they have learned about former routes during the learning phase. The approaches are perfectly capable of modeling proper routes inherited from the mathematical method relied upon a while learning. However, they also suffer from several drawbacks. The first one is shared with geofencing: there are plenty of possible routes between any two points, which are equally reasonable and are characterized by similar lengths and, therefore, costs. Nonetheless, if we compare such routes using a simple similarity measure, e.g. based on distances between trajectory points, routes that are almost equal become dissimilar. This phenomenon complicates the learning process and hinders the process of identifying a good model. The second problem consists in the verdict given by the algorithm (e.g. detection of a suspicious route) being difficult to explain to a supervisor. This is a known drawback of many learning methods, but it plays an important role in our application. Usually, the algorithm triggers a suspicious route alarm that is later taken over by an operator whose role is to either initiate verification, commence a pursuit or cancel the alarm. In order to make the decision, it is desirable that the operator knows why the system has triggered the alarm. Another drawback of learning methods consists in the fact that the learning phase has to take place, which means that the system cannot start operating without having collected a reliable history of routes,

possibly manually marked by an operator as valid or fraud prone. Such approaches may also be of the heavy-weight nature, both in terms of their conceptual complexity and computing resources required, especially during the learning phase. Selected authors are aware of these drawbacks and apply various measures to prevent them. For example, in [7], a one-class learning approach is used which simplifies the learning phase by making it human-independent. Addition of the two-class learning approach, when a human introduces negative (suspicious) examples, is possible as well. This optimizes the learning effect and also allows the learned model to follow human way of reasoning. The authors of [8] observe the phenomenon of multiplicity of equivalent routes. Thus they assess the current route not only by a usual point-distance similarity measure in reference to typical routes, but also by examining the total length of the connection. The expected distribution of the length for a given starting and ending point is calculated based on the historical files. Sadly, the authors only consider the total length of the vehicle's trajectory. They do not deal with the sections of fragments of the trajectory, potentially omitting some local deviations from the proper route. An interesting approach to online work and to the large scale of computing is proposed in [10]. As usual, a dissimilarity measure is used to compare the current trajectory with the historical ones. The potentially resource-heavy method is accelerated with special indexing of trajectories performed via the so-called local clustering and vantage trees.

In this paper, a fraud detection method based on on-line knowledge concerning development of the vehicle's trajectory is proposed. When the system detects an anomaly in the trajectory fragment observed so far, it triggers an alarm that draws the attention of the supervisor and requires them to take suitable action or to disregard the alarm. The proposed approach is based on a multi-resolution analysis of the route's economy, measured based on trajectory zigzagging (deflection from a straight line). This measure offers a multi-resolution functionality: both the shape of the entire route and the lengths of its fragments of various sizes are examined. The intention is to eliminate, from the algorithm, the many drawbacks referred to above. The approach accepts various route variants between the origin and destination locations and treats equally, as long as they exhibit similar economy metrics. It is conceptually easy and understandable for the operator assessing the alarm, who may easily assess the zigzagging of the route themselves. It is ready to run out of the box, without a lengthy and complicated learning phase, but some parameter tuning may still be necessary. The algorithm is also lightweight, facilitating its application in large-scale systems. It utilizes a para-stream character, and thus is suitable for on-line monitoring applications. Stream processing is strictly a synonym for "ideal data processing algorithm" [11]. Such an algorithm is characterized by constant memory usage per the size of input data. Thus, if the input data has the form of a stream of data, i.e. a set of GNSS frames for a given

vehicle, memory usage does not reach a certain limit as the stream proceeds.

2. The Algorithm

The algorithm described in this section is used for detecting anomalies in the economy of the observed moving trajectory. Normally, a vehicle should approach its destination using a straight and short route. Whenever this requirement is not met, the route is treated as suspicious. There is a possibility (risk) that deviation from the straight line was caused by illegal activity.

The concept of the algorithm may be easily explained by comparing it with geofencing. In geofencing, we define *corridors* along the projected route; whenever the corridor is left, an alarm is triggered. The most obvious drawback of this approach is the fact that multiple potential routes exist between the origin and the destination, as explained in the example of the grid layout of roads, as depicted in Fig. 1. Routes A, B, C look equally reasonable and the number of similarly reasonable routes grows exponentially along with the increase of the grid size. Therefore, it is not obvious how a specific corridor should be defined. If we define a "grid of corridors", i.e. those existing along all road segments present in the grid, we also accept such routes as D, which is lengthy and winding, and thus clearly suspicious in terms of potentially illegal loading and unloading operations.

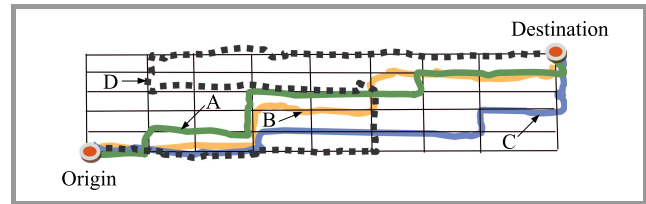


Fig. 1. Some potential routes between two points in a grid of roads. (See digital edition to find the color version).

Another drawback of corridors is that they do not allow any small deviations to rest areas, parking lots, tire shops, etc., especially if reaching these requires some maneuvers on one way roads and clover leaf interchanges.

Thus, in the proposed system, geofencing is replaced by an analysis of trajectory zigzagging. For given a trajectory T with points $x_1, x_2, \dots, x_n$ on a plane, we define its zigzagging $Z(T)$ as:

$$Z(T) = \frac{\sum_{i=1}^{n-1} \text{dist}(x_i, x_{i+1})}{\text{dist}(x_n, x_1)}, \quad (1)$$

where $\text{dist}(x, y)$ denotes the Euclidean distance between x and y . Alternatively, the length of the shortest path between x and y along the existing roads could be used for dist if map information is available, as would be discussed later. Zigzagging is simply the ratio between the length of the trajectory and the distance between its end points. The

main idea is to set a limit value ($Z^{\max}$) for the permitted zigzagging of vehicle trajectories. If trajectory zigzagging exceeds $Z^{\max}$, the trajectory becomes suspicious and illegal activity is taken into consideration. The rationale behind such a setting is that zigzagging routes suffer from bad economy – if the driver does need to go to illegal locations, they would not need to take such routes. Note, for example, that routes A, B, C (if represented as sufficiently dense discrete trajectories, i.e. trajectories with sufficiently large numbers of points) are characterized by a zigzagging measure of approx. the square root of 2. Route D has a larger zigzagging measure. This example shows that should $Z^{\max}$ be set at say 2 or a little more, in order to allow routing in a more complicated road grids.

Assessment of simple zigzagging measure values of the entire trajectory between the origin and destination points does not eliminate the potential of falsely evaluating the correctness of a given route based on zigzagging. It needs to be noted that if we take any fragment of the trajectory into consideration, it also should be economical in itself, i.e. should be characterized by a limited zigzagging measure. To cope with this, a *multiresolution analysis* of the route traveled so far, at a given point of the journey, should be performed. The zigzagging measure of the entire trajectory recorded so far, of a fragment of the trajectory ending at the current location, or even of a smaller fragment of the trajectory ending at the current location may be analyzed. Let us assume that the vehicle's route is given by the stream of points $x_0, x_1, x_2, \dots$, with x_0 being its origin. The proposed stream algorithm, invoked upon receiving frame x_i , is outlined as Algorithm 1 (variable *therewasanalarm* is initialized to false).

Algorithm 1. Response to a new trajectory point x_i

- 1: **If** *therewasanalarm* **Then**
- 2: **Return**
- 3: **End if**
- 4: Calculate the approximations of the subtrajectory ending at x_i :

$$T_1 = (x_1^1, x_2^1, \dots, x_{n_1}^1),$$

$$T_2 = (x_1^2, x_2^2, \dots, x_{n_2}^2),$$

...

$$T_r = (x_1^r, x_2^r, \dots, x_{n_r}^r),$$

where r is the resolution level, $n_1, n_2, \dots, n_r$ are not greater than the algorithm parameter N and $x_{n_1}^1 = x_{n_2}^2 = \dots = x_{n_r}^r = x_i$.

- 5: Compute $Z(T_1), Z(T_2), \dots, Z(T_r)$ (zigzagging for empty or one-point degenerate trajectories is assumed to be 0).
 - 6: **If** any of them exceeds $Z^{\max}$ **Then**
 - 7: Trigger alarm
 - 8: *therewasanalarm* = true
 - 9: **End if**
-

The proposed algorithm is also depicted in Figs. 2–3.

An alarm is generated only once for a given route in order to avoid the operator's distraction by a storm of alarms for consecutive trajectory points. This feature uses the Boolean *therewasanalarm* variable.

Note that since variables $n_1, n_2, \dots, n_r$ approximating sub-trajectories at particular resolution levels are limited by N , the number of resolution levels r is finite. Therefore, the algorithm conforms to the stream regime. The price that needs to be paid for this is that sub-trajectories may be approximated at N points at the most. In practice, it is difficult to implementing step 1 without storing all points $x_0, x_2, \dots, x_i$ in the memory.

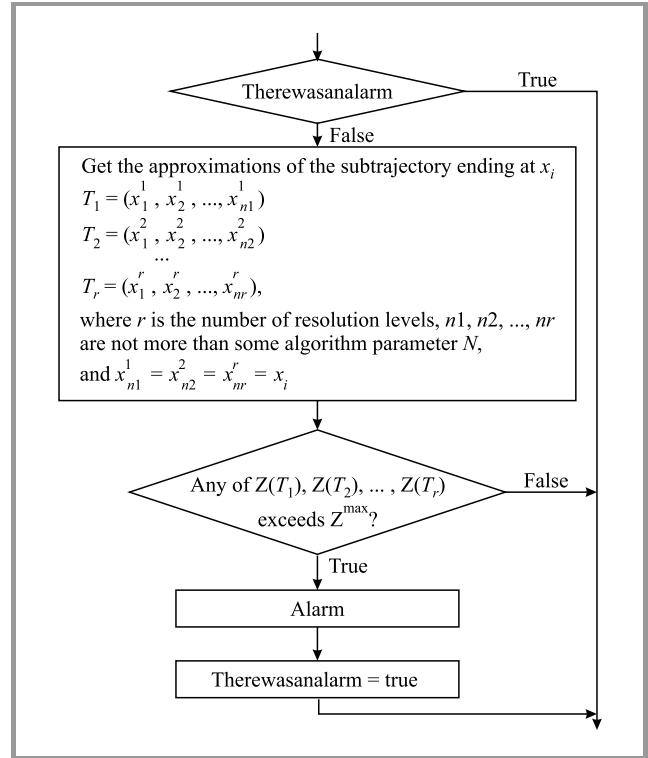


Fig. 2. Algorithm 1 used for processing a new trajectory point.

2.1. Selecting Route Approximations at Specified Resolution Levels

At each resolution level, the zigzagging of approximations $T_1, T_2, \dots, T_r$ of the sub-trajectory is examined. These approximations are some trajectories made up of selected points of the sub-trajectory that has been observed so far. The selection of points and the construction of T_i is based on keeping memory usage limited.

Determination of T_i is based on characterization of the resolution, performed by the reference length step for a given i . This reference length is called *stride_i*. This is given by:

$$stride_i = minstride^{2^i} \quad \text{for } i = 1, \dots, r.$$

The algorithm's *minstride* parameter defines the length of the reference segment of the trajectory at the finest resolution level. Parameter r itself is evaluated as:

$$r = \left\lceil \log_2 \frac{\text{maxroulength}}{\text{rrbuflen} \cdot \text{minstride}} \right\rceil,$$

where $\text{rrbuflen} = N - 1$ is the assumed maximum number of segments in T_i , maxroulength is the maximum possible length (sum of segment lengths) of the entire trajectory. The default parameter settings are: $\text{rrbuflen} = 4$, $\text{maxroulength} = 20,000$ km, $\text{minstride} = 2$ km.

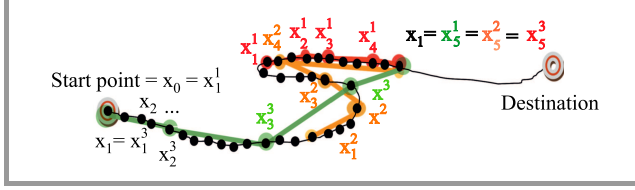


Fig. 3. Approximating the trajectory created so far with trajectories at 3 different resolution levels ($r = 3$). Probably, only the medium resolution trajectory t_2 (yellow) has a zigzagging measure that exceeds the alarm threshold.

Trajectories T_j for $j = 1, \dots, r$ are updated upon receiving a current point x_i of the trajectory according to Algorithm 2. Here, T'_j is a cyclic buffer storing a sequence of rrbuflen latest elements appended to this buffer. T'_j is initialized to empty sequence. Variable y_j is initialized to x_0 .

Algorithm 2. Update of trajectory approximation T_j with a new trajectory point x_i

- 1: $\text{stride}_i := \text{pow}(2i, \text{minstride})$
- 2: $r := \lceil \log_2 \text{maxroulength} / (\text{rrbuflen} \cdot \text{minstride}) \rceil$
- 3: **If** $\text{dist}(x_i, y_i) \geq \text{stride}_i$ **Then**
- 4: Append x_i to T_j (with deleting all but rrbuflen last element of the sequence stored in T'_j).
- 5: Set $y_i = x_i$.
- 6: **End if**
- 7: **Return** current T_j as sequence stored in T'_j (if it ends with x_i) or with appended point x_i (otherwise).

In other words, sequence T_j is designed by choosing points of sequence x_i . The first is x_0 , any new point x_i from the vehicle's trajectory is appended if and as soon as it is distanced from the previous appended point by at least stride_i . Such a sequence is appended the current point, if it does not end with it. At most last rrbuflen appended points are finally taken to form sequence T_j . Such an approach ensures limited memory usage during the construction process, which is $\mathcal{O}(\text{rrbuflen})$. Moreover, sequence T_j has segments of length of approximately stride_i , except for the last segment which was added to keep sequence T_j as fresh as possible.

The settings for r and stride_i (for $i = 1, \dots, r$) are selected so that:

- the logarithms of numbers stride_i form an arithmetic sequence,
- the coarsest resolution level, T_r is capable of covering the entire trajectory of the maximum length

maxroulength , i.e. to contain its first and last point. The T_i for other (i.e. those used to assess local zigzagging of finer resolution levels), might obviously be shorter.

2.2. Extensions

The algorithm is equipped with several extensions. Different zigzagging limits may be used for $Z(T_1)$, $Z(T_2), \dots, Z(T_r)$, instead of a single Z^{max} , in order to allow greater zigzagging at fine resolution levels. For given T_i , the limit is:

- $Z^{\text{max}, \text{short}}$ if $\text{stride}_i < \text{zigstridethreshold}$,
- $Z^{\text{max}, \text{long}}$ otherwise.

The reason for this extension is an observation that natural zigzagging of vehicles differs depends on the scale. In the case of routes that are less than 10 kilometers long, roads are often projected as being perpendicular to each other. Furthermore, drivers must perform maneuvers that increase the zigzagging measure, like bypassing one-way roads, making loops at multi-level intersections, obeying “no-turn” signs, etc. Hence, the default parameter settings are $Z^{\text{max}, \text{short}} = 2.0$, $Z^{\text{max}, \text{long}} = 1.7$, $\text{zigstridethreshold} = 10$ km.

The second extension allows to include the destination point in the trajectory that has been covered so far. If we know the trip's destination (which is usually the case), we may not only examine various zigzagging measures based on the sub-trajectory covered so far, but we may also proactively check whether various zigzags will need to become excessive in the future, assuming that the vehicle will eventually reach its destination. This extension works as follows. Upon receiving of consecutive point x_i , Algorithms 1 and 2 are invoked, but:

- the zigzagging excess check is performed as in Algorithm 1 (possibly with the extension of various zigzagging limits),
- the zigzagging check is performed again, not for approximate trajectories $T_1, \dots, T_r$, but for these trajectories, each with the destination point appended.

If any of these checks triggers an alarm, the entire algorithm generates an alert. This extension allows to trigger alarms in some situations. A typical situation is when a vehicle starts its trip in the direction that is opposite to its destination, in order to perform some illegal activity. Then, in the basic algorithm, any zigzagging activity observed increases extensively only after the illegal location has been reached and upon the vehicle starts moving towards the legal destination. However, an operator observing the route might be aware of the bad intentions of the driver even before the illegal location is reached. The vehicle driving in the opposite direction it allowed to continue doing so. By adding an artificial segment to the route covered so far, which predict the route ahead, allows the proposed algorithm to alert

the operator before the illegal location is reached. This advantage is illustrated by one of the experiments presented further on.

3. Discussion

Parameter *minstride* of the finest resolution naturally determines the rank of distance corresponding to “nonsensitivity level” of the algorithm to small deviations from the route. Along short sections of the trajectory, comparable with *minstride*, zigzagging may reach high values without triggering an alarm. This allows the driver to safely deviate to roadside facilities, such as hotels, restaurants or service stations. This mechanism allows for some other maneuvers as well: passing multi-level intersections, bypassing local one-way roads, performing maneuvers required by road signs. These trip disturbances may cause considerable local trajectory zigzagging, even imply trajectory looping, but as long as they are limited to a specific area of the *minstride* rank, they will be accepted by the algorithm.

Memory usage. Memory usage of the presented algorithm is $\mathcal{O}(rrbuflen \cdot r)$, where

$$r = \left\lceil \log_2 \left(\frac{maxroulength}{rrbuflen \cdot minstride} \right) \right\rceil. \quad (2)$$

The proposed algorithm assumes a limited length of the route and *rrbuflen* and *minstride* are arbitrarily chosen constants. While limiting the route length by constant *maxroulength*, constant memory usage is ensured. This assumption is not very restrictive in practice and memory usage increases only logarithmically with a potential increase in *maxroulength*.

Metrics used. In the algorithm, the length of shortest line on the Earth surface between two points x, y is computed by function $\text{dist}(x, y)$. However, there is a possibility of using map information instead, and of calculating the values of $\text{dist}(x, y)$ as the shortest road-based route between x and y . Some additional care must be taken to account for potential noncommutativity of such a distance. This increases the precision of the algorithm and boosts its independence from specific road network configurations. Certainly, the zigzagging limits would be decreased in such a scenario. However, such a solution is complex to implement in large-scale systems, i.e. for monitoring purposes. It requires additional memory to cache the map data locally in order to avoid extensive data traffic exchanges with the map server. Moreover, cache memory should also offer a mechanism for quickly calculating the distance between specific locations. Finding the shortest route online would clearly consume too much resources. The mechanism should perhaps contain some mesh of points with predefined distances and, based on this data, should approximate the distance between two given locations. This an interesting research problem in itself.

Traveling salesman case. It is often the case that a vehicle starting its trip has several destinations to visit. The algorithm may still be applied to such a scenario, even when the order of the visits is not known to the algorithm. The trajectory covered so far needs to be simply reset to empty and started anew upon any destination. This will allow only the zigzagging along the route section between two consecutive destination points to be examined. This fragment should be economical, i.e. no excessive zigzagging should be present. The extension expanding trajectories by destination locations is slightly more difficult to adopt in this case, since there are many destinations and we do not know which is to be reached first. Therefore, at the receiving point x_i , we should try all destinations and select the scenario that is most favorable for the driver. If no attempts trigger the alarm, the algorithm should keep quiet.

4. Numerical Demonstration

Several experiments with the proposed algorithm have been conducted. Their aim was to simply validate the presented approach, and to showcase the specific features and the behavior of the algorithm in typical scenarios. The extension providing various zigzagging limits was on, while the extension making use of the destination location was off by default, and was only switched on in one experiment. Default parameter settings were used.

A Java environment was used in the experiments. The code was written to retain the algorithm’s efficiency, i.e. cyclic buffers were used to collect the selected past points of the trajectory.

The routes were generated in using the Google MyMaps service as optimal routes form the town of Skierniewice to Warsaw (Poland). To obtain uneconomical variants of the route, specific mid-points were added to Google’s route finder. These points served as illegal locations, while the algorithm was still simply aware of Warsaw as the vehicle’s destination. The algorithm observed the trajectory online and triggered alarms, detecting suspected fraud. It needs to be remained that the algorithm is designed to alert once for each route to avoid a storm of alarms for consecutive trajectory points. The point of the trajectory at which the alarm occurred was registered and included in the subsequent presentation of the results.

The following experiments have been performed:

- the route from Skierniewice to Warsaw without any mid-points. It was expected that this route would not be assessed as suspicious;
- the route from Skierniewice to Warsaw with an additional mid-point in a forest in the village of Krakowiany, near the main road. This route includes an apparent deviation using a road leading to the forest and should trigger an alarm. The forest, however, is located near the main road (relatively to the length of the entire route). Thus, zigzagging measure along

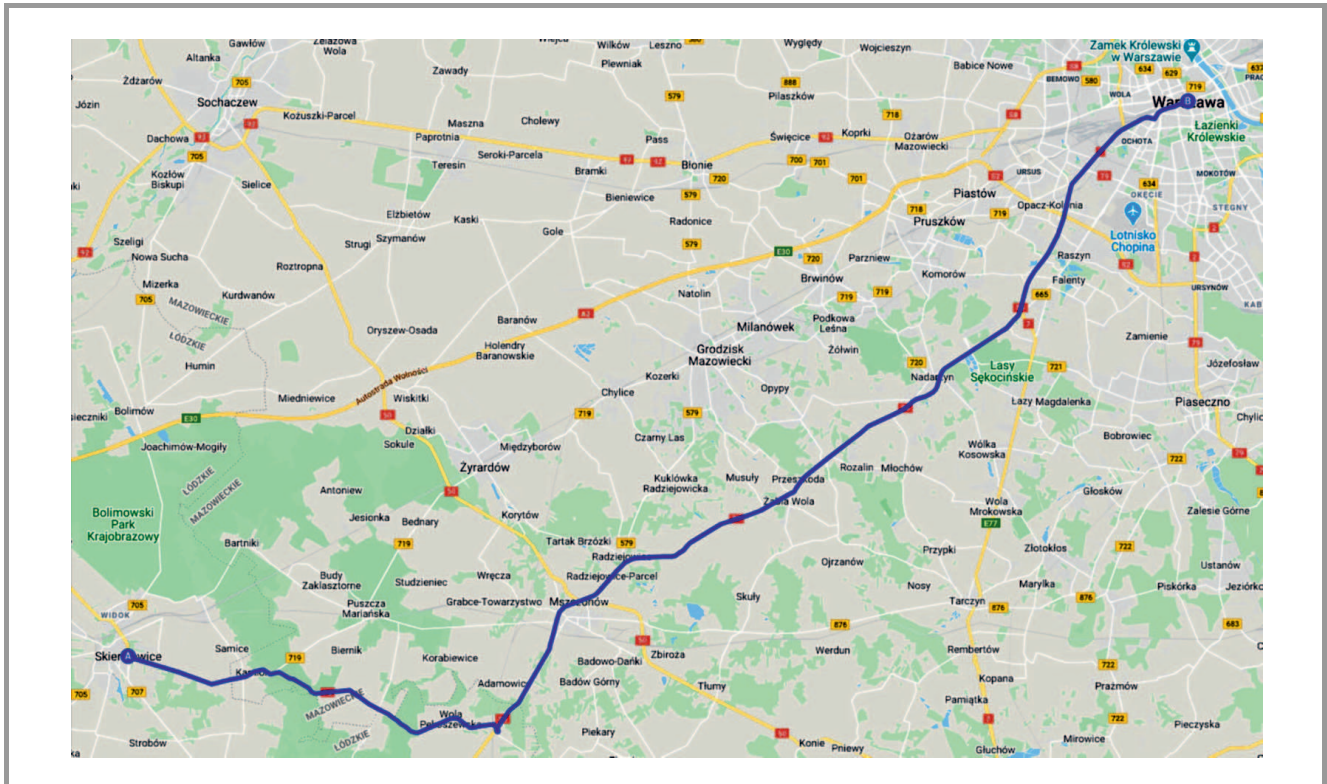


Fig. 4. Route from Skierniewice to Warsaw without a deviation – no alarm triggered.

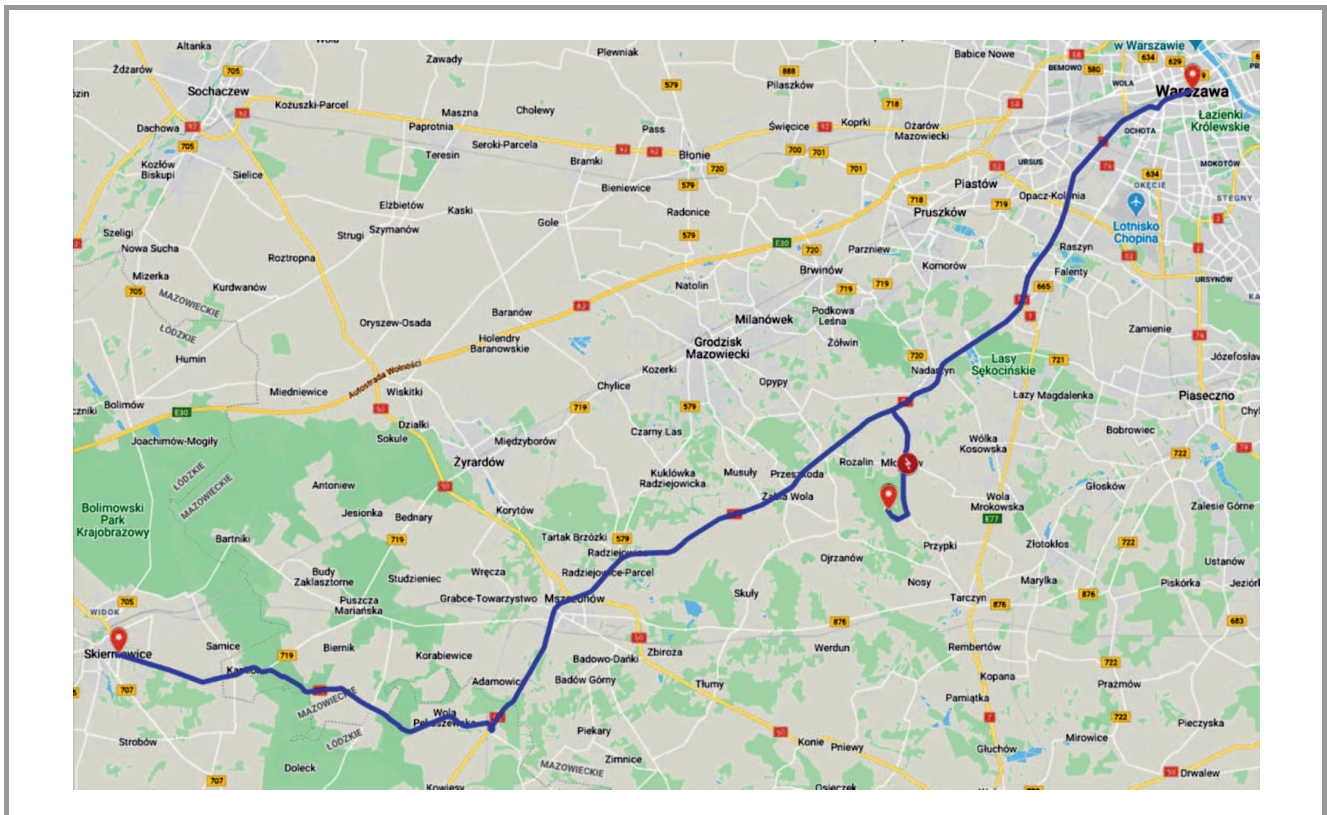


Fig. 5. Route from Skierniewice to Warsaw with a deviation to a forest in Krakowiany. An alarm has been triggered on the local road leading to Krakowiany, after visiting the village (during the return from the forest).

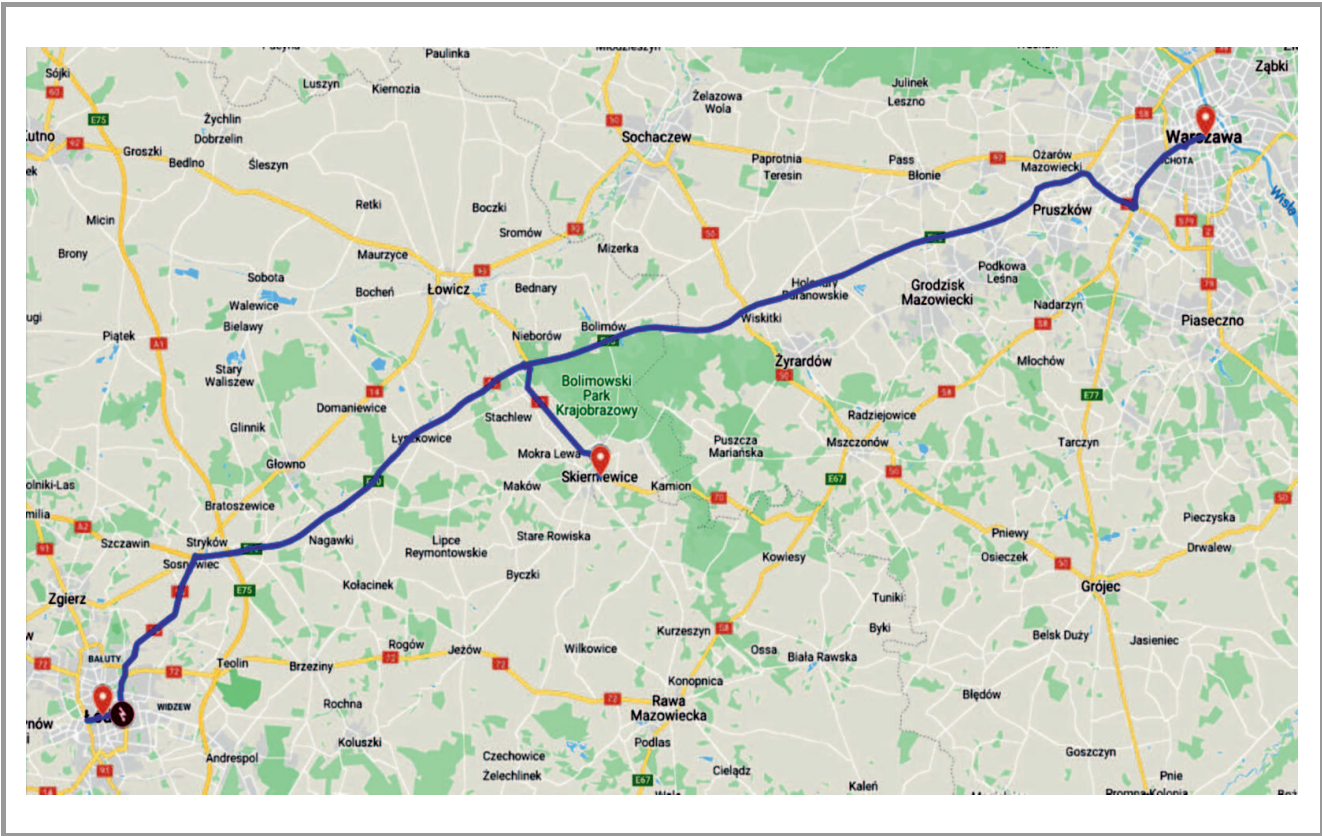


Fig. 6. Route from Skierniewice to Warsaw with a deviation to Łódź. The alarm has been triggered after visiting Łódź, on the return leg of the journey.

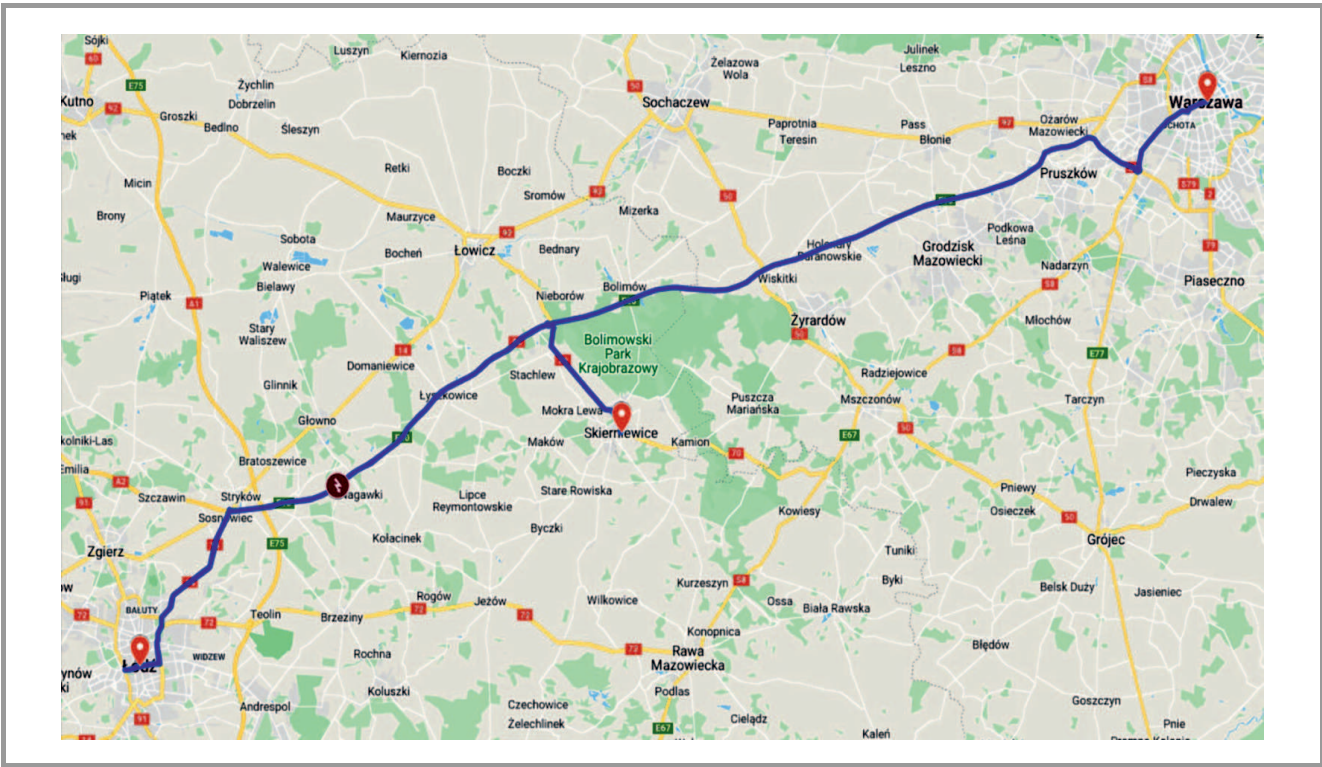


Fig. 7. Route from Skierniewice to Warsaw with a deviation to Łódź. The algorithm uses route augmentation, making use of the knowledge of the destination. The alarm was triggered before reaching Łódź.

the entire route could not be excessive. The example checks whether the algorithm reacts to an increase in zigzagging observed locally;

- the route from Skierniewice to Warsaw, with a mid-point of Łódź. When going to Łódź, a driver leaving Skierniewice needs to go in the opposite direction than when heading to Warsaw. Therefore, such a route should trigger an alarm;
- the route from Skierniewice to Warsaw, with a mid-point of Łódź and, unlike in the case of other experiments, the knowledge of the destination has been used in this test. This means that the algorithm augmented the trajectory covered so far with a segment between the current location and the destination. This augmentation should result in the alarm being triggered even earlier than in a scenario without augmentation.

The routes are presented in Figs. 4–7. The results of the experiments are presented in the form of icons in pink circles positioned at the locations where an alarm was detected. Since the roads on which the alerts were triggered are of the two-way variety, it is difficult to see, in the pictures, which direction the vehicle was driving in while causing the alarm. This is stated in the captions.

The alarms were triggered in concordance with our expectations. They appeared precisely along the routes with mid-points that emulated illegal load or drop locations. The route with the deviation to a forest shows the algorithm's ability to consider local route zigzagging. As far as the deviation to Łódź is concerned, the advantage of augmenting the route with a segment between the current location and the destination was shown. Without the augmentation, the alarm is triggered after visiting the town. With the augmentation, the alarm is generated earlier, on the way towards Łódź. This is beneficial, since the algorithm with augmentation is capable of making the same observations as a human being. The vehicle is supposed to head for Warsaw, but drives in the opposite direction instead. Also, the alarm is raised, in such a scenario, before the illegal activity takes place, which allows the proactive actions to be taken by the operator.

5. Conclusions

The proposed stream algorithm is capable of detecting truck fraud. In contrast to its numerous counterparts, the proposed solution uses an adequate and human-understood route economy measure. It is lightweight and does not require a learning phase. In experimental trials it turned out to behave reasonably and predictably.

References

- [1] S. S. Dukare, K. Rane, and D. A. Patil, "Vehicle tracking, monitoring and alerting system: a review", *Int. J. of Computer Applications*, vol. 119, no. 10, pp. 39–43, 2015 (DOI: 10.5120/21107-3835).
- [2] "PUESC" [Online]. Available: <http://puesc.gov.pl>
- [3] M. Saravanan, S. Aishwarya, and L. N. Aravindan, "Tracking anomalies in vehicle movements using mobile GIS", in *2013 Science and Information Conf.*, London, 2013, pp. 845–852 [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6661840>
- [4] A. Kuepper, U. Bareth, and B. Freese, "Geofencing and background tracking – the next features in LBSS" in *41. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2011 – Informatik schafft Communities, Berlin, Germany, October 4-7, 2011* [Online]. Available: <https://www.user.tu-berlin.de/komm/CD/paper/010221.pdf>
- [5] F. Reclus and K. Drouard, "Geofencing for fleet & freight management", in *2009 9th Int. Conf. on Intell. Transp. Syst. Telecommun., (ITST)*, 2009, pp. 353–356 (DOI: 10.1109/ITST.2009.5399328).
- [6] P. Deshmukh, A. Bhajbhakre, S. Gambhire, A. Channe, and N. Deshpande, "Survey of geofencing algorithms", *Int. J. of Comp. Science Engin. Techniques*, vol. 3, no. 2, 2018 (DOI: 10.29126/2455135x/IJCSE-V3I2P1).
- [7] R. R. Sillito and R. Fisher, "Semi-supervised learning for anomalous trajectory detection", in *Proc. of the British Machine Vision Conf.*, 2008, pp. 103.1–103.10 (DOI: 10.5244/C.22.103).
- [8] Y. Ge, H. Xiong, C. Liu, and Z. Zhou, "A Taxi Driving Fraud Detection System", in *11th IEEE Int. Conf. on Data Min.*, Vancouver, Canada, 2011, pp. 181–190 (DOI: 10.1109/ICDM.2011.18).
- [9] J. B. Oliv, "Anomaly detection and modeling of trajectories", M.Sc. thesis, CMU-CS-12-133, School of Comput. Sc., Carnegie Mellon University, Pittsburgh, 2012 [Online]. Available: <http://reports-archive.adm.cs.cmu.edu/anon/2012/CMU-CS-12-133.pdf>
- [10] Y. Bu, L. Chen, A. W. Fu, and D. Liu, "Efficient anomaly monitoring over moving object trajectory streams", in *Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discov. and Data Mining*, Paris, France, 2009 (DOI: 10.1145/1557019.1557043).
- [11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. London: MIT Press, 2001 (ISBN: 9780262032933).



Paweł M. Białoń is with the National Institute of Telecommunications. He received his Ph.D. in Automatic Control and Robotics from Warsaw University of Technology in 2013. His scientific interests include decision support and optimization with applications in telecommunications and data mining.

 <https://orcid.org/0000-0002-7781-9212>

E-mail: P.Bialon@il-pib.pl

Advanced Information Technologies Department
National Institute of Telecommunications
Szachowa 1
04-894 Warsaw, Poland

A Novel System Architecture for an Improved Self-care Solution – Conceptual Design and Key Components

Patrick Santos, Jorge Ramos, Eduardo Seabra, and José Castro

Uptec Parque da Ciencia e da Tecnologia da Universidade do Porto, Porto, Portugal

<https://doi.org/10.26636/jtit.2020.142920>

Abstract—The high penetration rate that mobile devices enjoy in to day's society has facilitated the creation of new digital services, with those offered by operators and content providers standing out. However, even this has failed to encourage consumers to express positive opinions on telecommunication services, especially when compared with other sectors. One of the main reasons of the mistrust shown is the low level of quality of customer service provided an area that generates high costs for the operators themselves, due to the high number of people employed at call centers in order to handle the volume of calls received. To face these challenges, operators launched self-care applications in order to provide customers with a tool that would allow them to autonomously manage the services they have subscribed. In this paper, we present an architecture that provides customized information to customers – a solution that is separate from mobile operating systems and communication technologies.

Keywords—*adaptive interfaces, big data analytics, data mining, user profiling.*

1. Introduction

Popularity that mobile devices, such as smartphones, enjoy in our society has led to the creation of new types of services available anywhere and anytime [1]. To get an idea of how important these devices have become, we may highlight the huge number of 2.4 billion smartphone users existing worldwide in 2017 [2] an increase of approximately five hundred million users compared to 2015 [3]. This is largely due to the variety of smartphones available on the market [4], and to the multiple services offered by content providers [5], [6]. However, despite the diversity of services that operators offer to their customers, their satisfaction with the telecommunications sector is often below average [7], and a notable number of complaints is reported each year by telecommunication sector regulators [8]. Clients refer to customer service as one of their most significant sources of dissatisfaction [9]. Customer service, is also one of the most significant cost items for operators themselves, due to the number of personnel needed to cope with the large volumes of calls received. In order to boost customer satisfaction, and also in order to lower the cost associated therewith, telecom operators

have introduced self-care applications [10], [11]. In general, self-care applications allow customers to use operators' digital channels to obtain information about their subscribed services, as well as to check account balance, the amount of mobile data used or, promotions offered by operators, to name just a few. As a result, access to customer service becomes more convenient for customers themselves, and operators may take advantage of a new channel to promote their services or products. This means that the potential for tightening ties between operators and customers increases [10].

However, despite all apparent advantages, mobile self-care applications suffer from certain issues that prevent their successful and large-scale use by consumers.

Unoptimized user experience. Many self-care applications made available by service providers suffer from inconsistencies in their navigation flow. For example, some users of Vodafone's self-care application MyVodafone [12] complain about numerous crashes that occur when the app is opening, and report inconsistencies encountered when taking advantage of existing features (e.g. problems with saving pdf invoices, no balance or amount of mobile data used shown). Surveys of self-care applications performed in other countries may be mentioned here as well, with the study described in [13], being one of the examples, where positive and negative aspects of Indian operators' self-care applications are pointed out. The features that affect user experience of these applications include lack of online customer support chats, failure to display information related to recent uploads, continued redirection to external web pages, or inability to add new numbers to the account.

Presentation of poorly customized content. This factor is intrinsically linked to low user experience levels. However, this problem is more far-reaching, since an appealing and well-constructed GUI does necessarily have to present content that is of interest to the user. According to some publications, self-care applications may actually help their users navigate through FAQ lists or surveys, but they provide access to limited, poorly personalized and artificial information [14]. In order to show that they really care about their customers, operators should provide customer support

chat tools within their self-care applications, as this is one of the communications channels preferred by users [15]. Other issues may be highlighted here as well, such as those related to the presentation of a few or no offers tailored to the user's profile, presentation of inappropriate value-added services, and scarce information about the services that have been subscribed to [13], [16].

Slow content delivery. Systems used to host all data and traffic are quite complex and often incompatible with today's needs and standards related to access to information. As the co-founder and vice president of Matrixx Software, a provider of digital services for the telecommunications industry, stated in [17], the problem with self-care systems, particularly in the case of large operators, is that they are moving away from real-time charging and billing systems. Thus, a great orchestration of efforts is required to update and retrieve data from, for example, 10 separate systems operating in the background. Thus, it is understandable why many users of self-care applications complain about data loading speeds experienced when using the online solutions [13], [18].

Misuse of customer data. Operators are able to access huge amounts of information about their subscribers. However, much of this information is misused, as it is relied upon to create differentiated services, including the provision of user-oriented products or services within the self-care applications themselves [19], [20]. In addition, customers' demographic and behavioral data may help create more personalized support mechanisms which, in turn, may increase the degree of satisfaction with operators' customer support services.

To address the issues related to self-care applications that have been mentioned above, the Adaptive Self-care project proposes to research and develop an innovative mobile self-care ecosystem enabling to provide information that is adapted to each customer's profile, via adaptive self-care interfaces. In addition, the ecosystem incorporates a data analysis and management tool to be used by telecom operators, allowing their managers to advertise services and promotional campaigns based on the customer's profile. Individual and anonymous profiles of each customer will be generated by robust big data analytics engines.

In addition to presenting the innovative mobile self-care ecosystem referred to above, this paper also describes the related work related to mobile operators' self-care applications (Section 2), the research methodology used to better define system requirements that need to be complied with (Section 3), and the resulting system architecture (Section 4). The paper is concluded with a discussion about the results obtained and with presentation of potential future work (Section 5).

2. Related Work

With the Adaptive Self-care solution presented in this paper, we aim to offer telecommunications operators, as well as their respective customers, an integrated, cloud-based

solution that complies with the latest security certifications and standards, and contains mechanisms for big data analytics, machine learning, deep learning, AI, as well as data monetization and governance. The said solution aims to provide operators (via a web application) and customers (via a mobile app) with a number of advantages, such as cost reduction for operators and increased efficiency and customer satisfaction levels for users. However, it needs to be borne in mind that similar mobile self-care solutions, as well as data analysis and management tools are already available on the market. Therefore, this chapter provides a summary of some of those solutions.

The solution offered by **Vodafone** [12], one of the largest telecommunications operators in Portugal, consists of a web tool and a mobile application adaptable to different screen resolutions of mobile devices, such as smartphones and tablets. It is available for Android and iOS, uses GPS, GSM/4G connectivity and Wi-Fi to enable customers to manage their accounts and services. The app is available in three languages and allows customers to control data consumption, calls and SMS, fill up their accounts and make payment, view invoices, access technical support, receive promotional offers, manage loyalty programs and services, control data and roaming settings and discover all available service. However, personalized content is provided on a very limited scale only.

MyBusiness [21], [22] is an app intended for business customers of Orange, enabling them to manage their companies' telecommunications accounts and data consumption. The app offers the following features: management of all Internet connections, landlines, mobile accounts and data consumption – from one place at any time, management of the company's overall balance, management of account refills and payments, as well as management of multiple accounts and services assigned to specific employees/devices. Despite the features offered, this mobile app is only available for business users and lacks content personalization.

Mitt Telenor [23], Norwegian version of the My Telenor application developed by Telenor Norge, is a mobile customer service application designed by Telenor. The app allows to perform the following activities: buying additional data volumes, managing payments (including the ability to postpone them), managing service subscriptions, locating the nearest Telenor store, making pay-as-you-go payment via Vipps, and accessing customer support services, including SIM card ordering and management, as well as accessing PIN and PUK codes. However, consumers are not provided with access to targeted campaigns advertising products of their interest, which indicates that consumers information is not used to full extent.

The solution adopted by Portuguese telecommunications operator **NOS** [24], [25] allows customers to manage television, Internet, mobile phone and landline services, namely: balance and data consumption, tariff changes, shipments, access to invoice details from the past six months, access to PIN and PUK codes, as well as access to specific applications of each service. However, nothing is said about

personalization of content that is provided to customers, e.g. about providing access to targeted marketing campaigns. In addition to mobile self-care applications for telecom customers, we have also identified data management tools and big data analytics platforms available on the market for telecom operators.

Nokia Telecom Analytics [26] is a Nokia solution for big data analytics targeted for telecom operators. It incorporates cognitive data analysis (Nokia's cognitive analytics) methods based on cloud machine deep-learning algorithms that enable real-time monitoring of the data infrastructure, network and connected devices for process optimization and automation, as well as anomaly prevention, and allows to provide optimized customer care services. Despite providing a quality monitoring service, this platform does not make use of consumer data to improve the level of marketing services offered by telecom operators.

Mahindra Comviva's mBAS [27] is a Mahindra Comviva solution, widely used on the Indian market. It enables real-time data monetization and offers a distinctive feature that allows to pay for the services provided by the operator or by third parties using the customer's current balance – the solution is known as direct carrier billing and enables operators to increase revenue by collaborating with third-party digital content providers, such as YouTube and Spotify. It also allows the configuration of generic services and products through a unified platform. No additional information is available on how it uses data provided by consumers to create better offers (services, products, tariffs, etc.).

Calero [28] is an integrated communications lifecycle management (CLM) platform covering landline and mobile lines, as well as mobile data transmission (TDM, MPLS, VoIP and wireless). It aims to provide the operator with a business intelligence solution relying on data analysis to optimize profitability of their business. It automates the process of issuing paper or electronic invoices, facilitates payment processing, including charge allocation, allows to approve payments, traces usage-based internal recoveries, offers detailed multi-currency settlements, audit functions, dispute resolution features (detection of billing errors, downtime, potentially fraudulent activity, misuse/abuse), and provides customer support features, controls the distribution and sharing of services, software licenses and digital assets.

Comarch Telecommunications CEM is Comarch's solution [29] for managing and analyzing the quality of services provided by telecom operators, with a particular focus on increasing customer satisfaction. One of the key requirements for optimizing the experience and the resulting customer satisfaction level of telecom operators is uninterrupted availability of services. Therefore, this solution incorporates an automated fault management tool. This network failure management tool has been adopted by such operators as Vodafone, T-Mobile Poland, Telefonica O2 Germany and MTS Russia, and may be expanded using topological correlation and machine learning techniques to ensure more efficient management of and to prevent net-

work failure alarms. The CEM solution integrates the fault management tool with OSS/BSS big data analytics capabilities to provide the operator with real-time monitoring not only of network performance but also of the customers' perception of the network, allowing the operator to take preventive measures in order to guarantee the provision of services and to prevent network failures.

Development of services that are more adequate to consumer preferences in the telecom area is also a research topic for the scientific community. For example, in [30] a study is presented of a new customer segmentation mechanism for the mobile telecommunication sector, based on service usage behaviors. This work relies on data mining technologies deployed for customer clustering and segmentation based on their detail call records and data consumption behaviors. The obtained results are a source of new knowledge about the impact that consumer behavior exerts on their loyalty and satisfaction levels. Other works, such as [31] and [32], also discuss the implementation of advanced techniques for customer segmentation in order to better understand their needs. Nonetheless, neither the scientific community nor the commercial telecom sector have identified a solution that would be capable of both satisfying consumers and telecom operators, since there no standard technique to acquire data concerning consumers in order to optimize the services provided to them has been defined, and no tool or framework capable of monitoring all aspects related to consumers and the network exists as well, which would also allow all services provided by telecom operators to be optimized. Therefore, an intelligent big data management and analytics tool intended for operators, once developed, may prove to be a highly innovative solution for the telecom sector.

3. Research Methodology

For the purpose of this project we are using the action-based research methodology that is iterative and user-centered. A session in which customers of telecom operators participated as the target audience for mobile applications that integrate the adaptive self-care ecosystem has been held. The session took place in three stages: presentation, synergistic debate among the participants and completion of a questionnaire required to conduct a statistical survey concerned with functional requirements and eventual specificities of the mobile application that need to be taken into consideration in the Adaptive Self-care ecosystem. The presentation aimed to introduce the concept of self-care relied upon in the telecommunication operators' systems, which is the main focus of the project. It also served as a point of departure for a synergistic debate, facilitating the alignment of customer interests with the general objectives of the project. Finally, completion of the questionnaire allowed to document, in more detail, the aspects discussed, in order to obtain statistic results allowing to more precisely define the app's priority features.

Given the statistical results and suggestions obtained via the survey, it was concluded that the session with the participation of customers of telecom operators was beneficial for the project. In addition to proving the clients' demand for such an innovation, it was also possible to set the functional priorities of the application that will contribute, from the end users' point of view, to a higher level of usefulness, flexibility and easier adoption of the system. The analysis of requirements took into account all functionalities proposed in the course of the session, including those considered appropriate and feasible, and enabled to properly adjust those that have already been foreseen in the basic assumptions, to make sure that they comply with, the legal requirements and fit in the context of the project. The goal was to include those features that boost the overall potential of the original project, and not just those that were suggested.

After a comparative analysis of the functionalities offered by commercially available solutions has been performed, and after the study with the individual participants was conducted, it was possible to identify some of the desired functionalities for the Adaptive Self-care solution. This means that both functional and non-functional requirements of the platform could be defined.

3.1. Functional Requirements

Functional requirements include those that describe the behavior of the application, i.e. functions or services that the application must be capable of performing. The list below presents the functional requirements applicable to mobile applications intended for customers of telecom operators:

- the application should allow the client to analyze and manage data via lists and graphs, with the filtering option available,
- the application must offer the ability to connect to the Internet via Wi-Fi or mobile data,
- the application should be able to collect geographic coordinates using the smartphone's GPS, in order to present location-based content,
- the user must have the ability to add multiple devices to the same account,
- the user must be able to add several accounts to his/her profile (e.g.: 1st party household with Internet, TV, landline, and cell phones of his/her own and their dependents; dependents in a 3rd party household),
- the user should be able to register in the system and edit his/her profile later,
- the application must provide notifications of events/promotions/payment deadlines,
- the user must be able to manage privacy policies,
- the user should be able to locate stores of the operator and of the business partners, using the built-in map,
- the application should provide access to customer service through a chatbot,
- the application must allow to search the operator's products/services,
- the application should allow to subscribe to new services and to purchase products/add-ons.

3.2. Non-Functional Requirements

Non-functional requirements are those that relate to application's performance in terms of usability, reliability, security and accuracy of the services offered. The non-functional requirements of the Adaptive Self-care mobile application include the following:

- the application interface should be simple and objective,
- the application interface should be flexible, in order to adapt to different resolutions and screen sizes of devices using both Android and iOS operating systems,
- user data should be stored remotely (cloud) and anonymized in order to ensure the privacy of users and of their dependents and, with their consent, the usage of that data by the operator's big data analytics engines which, by its turn, should be capable of enhancing personalized services and promotions,
- the application should be able to submit request responses within 10 seconds. Otherwise, a timeout message should be displayed,
- the database in which non-anonymized user data will be stored should be password protected to prevent unauthorized access,
- Android and iOS applications should be compatible with AI, machine learning, deep learning and big data analytics algorithms operating in the cloud environment,
- compliance with data security and privacy standards applicable to telecommunication services must be ensured.

4. Results

Having analyzed different self-care solutions, we propose a cloud infrastructure that satisfies the needs and fills the gaps detected. The central block of the entire system is based on cloud infrastructure. All processing, security and information storage operations are performed here, and the connection between the different applications and external

systems is ensured. The central block consists of different web services that support mobile applications and backoffice platforms. The cloud infrastructure is also used for creating notifications, collecting and aggregating statistics and routines for consumer profiling, as well as for performing operations that validate business flows. More specifically, the cloud infrastructure comprises the following elements:

- NGINX load balancer [33] which receives and manages all requests, efficiently balancing load and evenly distributing requests across available servers,
- Tomcat server [50] which acts as an application container and allows the implementation and deployment of applications developed using the Java technology,
- MySQL [35], as a database that contains all information stored in a structured way,
- Redis [36], as a non-relational database and data cache,
- Cron jobs which allows tasks to be automatically and periodically scheduled at a pre-set date and time,
- logging system which allows the logging system to be integrated easily, quickly and neatly,
- RESTful web services [37] using the JWT standard for tokenization [38],
- Swagger, as an API development support framework.

The cloud infrastructure of the Adaptive Self-care ecosystem needs to support different types of services, such as storage, business flow processing, business intelligence mechanisms, and smart communication (notification) methodologies. The cloud infrastructure allows the aforementioned services to be integrated, aggregated, customized and delivered to end users based on a distributed and on-demand strategy [39].

4.1. Architecture

The cloud infrastructure needs to respond to a number of challenges in order to successfully support the proposed features. Thus, a scalable system needs to be implemented due to the large amount of information that needs to be stored and processed. The system also has to provide forms of communication for the mobile and web applications, including security mechanisms, operator service information processing, consumer profiling algorithm processing, and automated information synchronization tasks.

In order to create an ecosystem that comprises all required components, it has been based on a service-oriented architecture (SOA) model [40]. The SOA architecture model allows to integrate a variety of services that may be exposed or not, depending on the consumer's needs. It also allows to share the services and enables them to communicate with each other. This model also offers flexible integration and cloud web services reusability through

a service-based modular architecture, encapsulating several applications and data sources in the form of a black box [41], [42]. The services are exposed based on the RESTful web services [37] standard, i.e. one of the most widely used specifications. REST (Representation State Transfer) is a simpler form of communication than traditional SOAP [43], since the information sent is less detailed and overly complex communications are avoided.

Information is sent in the JSON format [44] that requires less data processing and enables easier interconnection between the services. It is a format-independent technology, since each platform already has its own data interpretation mechanisms. The services are exposed externally through application program interfaces (APIs). APIs are interfaces that enable users to consume cloud services. Development performed outside the APIs is considered part of the software itself, i.e. customer-exposed software [45]. Thus, the project architecture is based on the Software as a Service (SaaS) solution [46]. In SaaS, software is distributed as a service and is implemented natively on one or more web servers. The main goal is to deliver application services to the users as per their requirements [47]. An Internet connection is required to use the services exposed by the software. From the architectural point of view, one may say that SaaS guarantees the following advantages:

- **scalable infrastructure** – cloud resources may be expanded quickly and cost-effectively, without disrupting the services,
- **availability** – in accordance with information specified under the previous item, the services are available and accessible anywhere and anytime, thus contributing to greater mobility,
- **time and cost savings** – it is possible to make the services available to new customers using the same infrastructure, with only a few initial configurations. Such an approach shortens the lead time required to launch the services to new customers and reduces technical costs,
- **new features** – it is possible to release new features and improvements according to the platform's needs.

4.2. Information Processing

In this solution, information is processed in the cloud. Thus, applications (mobile and backoffice) avoid processing huge quantities of data locally. Such an approach is consistent with the line of thought relied upon at all stages of the project development process and allows to avoid various scenarios:

- **Slow data processing on the application side.** Avoiding complex operations on the users' devices allows to eliminate unexpected errors and ensures smooth operation of the application. If information is processed slowly in the cloud, it is possible to analyze the problem and avoid changing applications. If

it is a hardware resource issue, cloud-side resource scalability is quite easy to perform due to the nature of the architecture implemented;

- **Extreme battery drain.** Mobile device battery usage must be taken into account while performing complex processes;
- **Security flaws.** Communication between applications and the cloud is secure due to the implementation of HTTPS certificates. Thus, sensitive data are only available in the cloud and are not passed outside;
- **Bad project development practices.** To avoid this issue, mobile/web application developers and back-end programmers must observe good project development practices.

4.3. Data Mining

The platform uses data mining mechanisms to find useful usage patterns when analyzing data collected by the system. These patterns are relied upon to create user profiles and to automatically adapt their interfaces. Data mining algorithms were applied to acquire knowledge that is based on the data collected, to filter relevant information on the quality of offers and/or discounts, as well as to perform automatic or semi-automatic cataloging and targeting of offers and/or discounts according based on client segmentation. The process of tailoring interfaces to customers must primarily focus on service customization aspects and on quick access to the solution's functionalities. Data mining mechanisms focus on multiple types of customers' behavioral and/or personal data, and they aim to identify information that is most relevant to building consumer profiles and to tailoring the content accordingly.

To achieve the intended objective, the alternating least square (ALS) algorithm was chosen. It is, one of the algorithms that are most frequently used in recommendation systems, and is better known as collaborative filtering (CF). To run the selected algorithm, we used the Spark framework. The intended purpose is to associate the user profile with segments pre-created on the platform (segments may refer, for instance, to product types or tariff types). Information concerning product visualizations and purchases was used. The algorithm relying on this type of data defines product preferences and product types for each consumer and runs through weekly routines to update the results. When the routine commences, the algorithm uses a percentage of the data to create a training file and the remaining percentage to obtain the final output. Once the algorithm is finished, the information is stored in the Redis cache, allowing it to be updated by each user.

When the user checks the suggested products in the mobile application (self-care mobile app), the information is ready to be consumed through the REST web service. To perform the mining task, information about the consumers, i.e. about their behaviors (products that they searched for)

and about their purchases is used. A mono-database mining strategy (mining is performed on data sources to identify global interest patterns) was used, as indicated in [48], since the number of data sources needed for the mining task was not high. The main goal of this strategy was to build a recommendation system for clients that would use the self-care mobile app developed under the project. To achieve that purpose, the first step was to group all consumers with similar preferences by using the k-means algorithm ($k=3$) and Weka software. Figure 1 presents an example of the data used in the study, containing information about viewed and purchased products. The following fields were taken into consideration: consumer ID, date, time, product ID and product type.

```
(...)
13, "2019-04-13", "14:55:29", 1, 1, "Smartphones"
32, "2019-10-06", "18:08:21", 15, 1, "Smartphones"
25, "2019-03-02", "04:45:22", 2, 2, "Accessories"
70, "2019-02-01", "08:20:46", 3, 3, "Tablets"
100, "2019-09-23", "23:38:21", 15, 1, "Smartphones"
2, "2019-05-11", "00:32:04", 4, 2, "Accessories"
43, "2019-01-01", "14:39:27", 25, 2, "Accessories"
(...)
```

Fig. 1. Example of data used to group all consumers with similar preferences.

Then, clusters were created by using the k-means algorithm. Cluster 0, with 31% of total data (91 instances), is related to customers who viewed/purchased products within the tablet category. Cluster 1 (47% of total data – 136 instances) is related to consumers who prefer smartphones. Cluster 2 (22% of total data – 62 instances) is related to customers whose preferences focus on accessories. Figure 2 depicts the result instances for every cluster. These clusters grouped all consumers based on their purchase preferences. Thus, clients of each segment (cluster) would receive information about campaigns promoting tablets, smartphones or accessories, respectively.

Attribute	Clusters			
	Full data	Cluster 0	Cluster 1	Cluster 2
	286	91	136	62
Date	2019-02-08	2019-02-08	2019-03-26	2019-03-26
Time	10:46:17	10:46:17	16:45:07	15:30:26
Product	1	6	1	4
Product type	1	3	1	2
Class	Smartphones	Tables	Smartphones	Accessories

Fig. 2. Clusters created after applying the k-means algorithm.

4.4. Big Data Analytics

The prototype aimed to use all information originating from users in order to create a differentiated platform. The main

concern related to the data structure itself and to the technologies used. In practice, a machine learning tool Apache Spark [42] was used, which basically processes information stored in databases, generates training files (csv), and then processes the saved data output. The framework handles and builds files from resilient distributed datasets (RDD's).

These RDDs may contain any object which operates as a capsule around a small or large dataset, allowing the creation, transformation and invocation of different types of operations within them. The Apache Spark Mlib library allows the usage of operations and algorithms offering fast and high-quality performance parameters in the Java technology, therefore guaranteeing results that are better than those achieved by other solutions available on the market. During this project, different machine learning algorithms were studied, such as: tree methods, logistic regression, linear regression, recommender systems, to name a few, with the ALS recommender system algorithm proving to be the best choice for implementing the project's artificial intelligence solution.

ALS is basically a matrix factorization approach that implements a recommendation algorithm which may be used on e-commerce platforms to recommend products to customers. Unlike some client-based or product-based recommendation systems that calculate similarity, the ALS algorithm uses weight factors, returning a final and justifiable rating from the performed study. In our case, together with the machine learning algorithm, two files were created. The training file will be used, as the name implies, to train the algorithm to obtain more accurate results. The test file, in turn, will return the desired outputs to their APIs. Both files work as a three-phase feedback system. During the first phase, the desired information is collected from the dataset from the training file. In the second phase, where the algorithm uses both files. Relying on the training file, it analyzes customer preferences, based on such attributes as customer ID, the product ID, number of views and purchases made in connection with a given product. Based on that data, the algorithm calculates a percentage ratio and moves on to the next phase. In the third phase, the information is sent to the test file, where it fetches the desired output, either for a particular product or a product type. These files are created in a separate project, where in turn they are stored in a dev environment and updated weekly with customer's actions. For this there are two jobs, allocated in cron job `MIClientSegmentProduct`, which binds clients to segments and `MIRecommendedProduct`, which creates the customer recommended product files. In short, the entire process allows customers to be associated with different types of segments, according to their preferences, and to consequently receive product recommendations that are potentially matched to their preferences. A caching system, Redis, has also been implemented to make sure that, in situations where the customer wishes to consult the recommended products, it is not necessary to run the algorithm anew, also making this process faster. A big data analytics mechanism based on a local pattern analytics strategy, such

as the one demonstrated in [49], may be used for developing system prototypes, as it may decrease the complexity of the algorithms used to find relevant patterns and may avoid potential data ingestion-related issues.

4.5. Push Notifications

As the popularity of smartphones increases, push notifications have become an easy methods used to immediately send information directly to the user's device. Not only do they ensure alert visibility, but they are an important method for maintaining and strengthening customer leads, making them an attractive feature for deploying specific marketing strategies [50], [51]. In Adaptive Self-care, notifications are sent using the Firebase Cloud Messaging (FCM) platform [52]. FCM is one of the components of Google's mobile development platform and supports iOS, Android and web clients. In addition to push notifications, it also sends data messages to be processed by the application. Messages may be intended for three types of targets: a single user, groups of users, or users who have subscribed to specific topics using a given service. As part of the project, several types of push notifications were implemented to provide users with relevant information only. The following are the scenarios adopted for each notification type:

- Limits concerning data, SMS and/or voice usage. The tariffs made available by the operator allow users to enjoy data, voice and SMS. As a rule, each component described above has a predefined limit. Thus, whenever the user exceeds 80% or 100% of their limit value, he/she receives a notification. The first notification helps the user get an idea of how much they have spent in the current month. Once the user reaches 100%, they are notified again, and may their access to a given activity/component (data, voice, SMS) may be prevented;
- Generation of a new payment. Users who are post-paid subscribers receive, when an invoice is generated, receive a notification to let them know they have a payment to make. Pre-paid users receive a notification informing them that a given amount has been deducted from their balance or that they have to refill their account to continue enjoying the service;
- New campaign/product. The operator's employees who upload campaigns may trigger a new notification for the system. Using the "create new campaign" form, backoffice users have the option to select segments to which the notification will be sent. Segments are created automatically by the AI module based on the machine learning algorithms that associate users with those segments. This mechanism enables the platform to send targeted notifications to users based on customer profiling;
- Notification on the status of an open report. Using the app, customers may report a problem with their

service or a bug found in the application. When the operator's employee processes such a report, they can send a notification of the current status of the issue.

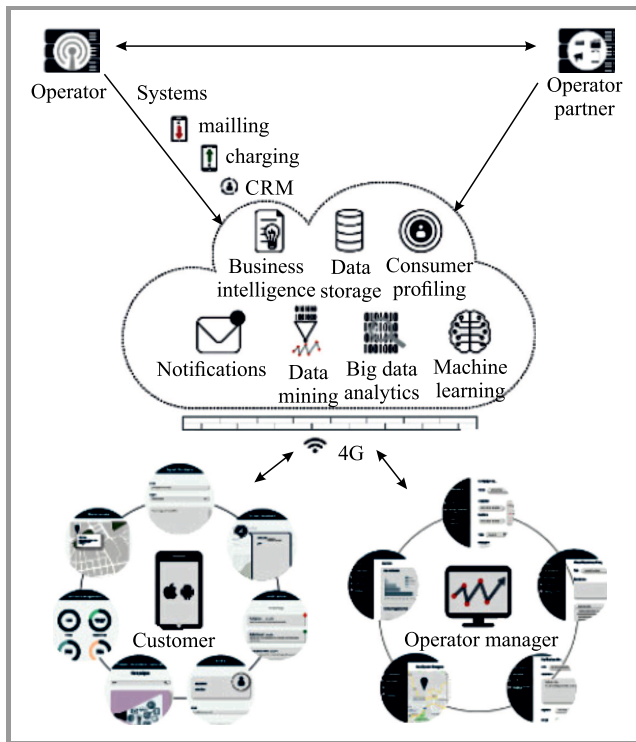


Fig. 3. High-level system architecture.

The notifications that have been sent are stored in the database and their status lets the users know if they have actually been sent or not. When the connection is not available or a service-related problem occurs while sending the notification, a procedure is applied automatically that re-sends the notification. All data, such as the message body, the content to which it refers (if it is related to campaign, product or payment), the creation date, the actual sending date and the indication whether the user has read it, is stored. Having all notification-related information stored allows the operator to understand whether the notification module is working properly. An API is also available via the web service, allowing users to get all notifications they have received. Such a solution provides the history of notifications via the mobile application. Notifications that have been read are distinguished from ones that remain unread, and their status is marked in the app itself (notification badge).

5. Conclusion and Future Work

Our study found gaps in the mobile self-care solutions actually offered by telecom operators. The business management and analytical platforms used by telecom operators also fail to make the best use of consumer data available to them, wasting the opportunity to deliver optimized services and products to consumers. Based on this study, the

requirements for the Adaptive Self-care solution were defined, enabling the construction of a cloud architecture that meets those requirements pertaining to storage, processing, scalability and functionalities. For future work, we will do three iterations of the action research methodology, comprising the phases of research, development of the prototype and evaluation (which serves as a point of departure for the next iteration). During the last iteration, the final prototype will be deployed and tested with a specific group of telecom operator customers, as an operational pilot, for the period of one month.

6. Acknowledgement

This work is a part of the Adaptive Self-care project, co-funded under COMPETE/P2020/EU, within the framework of the Portuguese Sistema de Incentivos à IDT Empresarial (project 038259).

References

- [1] D. Lamey, "The evolution of technology: past, present and future", *Discover Tech*, 2018 [Online]. Available: <http://www.discovertec.com/blog/the-evolution-of-technology> (accessed on 2019-11-02)
- [2] D. Murphy, "2.4 bn smartphone users in 2017, says eMarketer", *Mobile Marketing Magazine*, 2017 [Online]. Available: <https://mobilemarketingmagazine.com/24bn-smartphone-users-in-2017-says-emarketer> (accessed on 2019-01-03)
- [3] "Number of smartphone users worldwide from 2014 to 2020 (in millions)", *Statista* [Online]. Available: <https://www.statista.com/statistics/330695/number-of-smartphoneusers-worldwide/>
- [4] C. Low, "It's no longer enough for budget phones to just be cheap", *CES 2017*, 2017 [Online]. Available: <https://www.engadget.com/2017-01-08-budget-phones-ces-2017.html> (accessed on 2019-10-04)
- [5] "Condições de oferta e de utilização de serviços de comunicações eletrónicas da MEO" [Online]. Available: <https://conteudos.meo.pt/meo/documentos/condicoes-oferta-servicos/mod-c1001276.pdf> (accessed on 2019-10-25) [in Portuguese]
- [6] D. Nilsson, "Examples of value added services VAS in telecom 2017", *Appland*, 2017 [Online]. Available: <https://blog.applandinc.com/examples-mobile-vas-services-vas-telecom-2017> (accessed on 2019-02-14)
- [7] "Benchmarks by sector", *ACSI* [Online]. Available: <https://www.theacsi.org/acsi-benchmarks/benchmarks-by-sector> (accessed on 2019-10-17)
- [8] D. de Notícias, "Queixas nas telecomunicacoes aumentam para 69 mil", 2017 [Online]. Available: <https://www.dn.pt/sociedade/interior/queixas-na-area-das-telecomunicacoes-invertem-tendencia-e-aumentam-para-69-mil-em-2016-7205687.html> (accessed on 2019-01-04) [in Portuguese]
- [9] K. Clarke, "Why mobile operators must break free from bad customer service", *Telecom Tech News*, 2017 [Online]. Available: <https://www.telecomstechnews.com/news/2017/jul/28/why-mobileoperators-must-break-free-bad-customer-service/> (accessed on 2019-01-05)
- [10] "Mobile self care apps – customer support benefits for telecoms", 2016 [Online]. Available: <https://www.kochartech.com/mobile-self-care-apps-customer-support-benefits-telecoms/> (accessed on 2019-05-05)
- [11] "Why seamless self-care should be a top priority for telcos?", *Sterlite Tech*, 2017 [Online]. Available: <https://www.stl.tech/blog/why-seamless-self-care-should-be-a-top-priority-for-telcos/> (accessed on 2019-05-05)

- [12] "My Vodafone Play store web page", 2018 [Online]. Available: <https://play.google.com/store/apps/details?id=com.myvodafoneapp> (accessed on 2019-02-05)
- [13] "Comparison of self-care apps offered by Airtel, Vodafone, Reliance, Rircel, Tata Docomo and BSNL", *Telecomtalk*, 2016 [Online]. Available: <https://telecomtalk.info/comparison-of-selfcare-apps-offered-by-airtel-vodafone-reliance-aircel-tata-docomoand-bsnl/139985/> (accessed on 2019-08-22)
- [14] V. Narang, "Telco Dilemma: Prepaid vs Postpaid", *LinkedIn*, 2015 [Online]. Available: <https://www.linkedin.com/pulse/telco-dilemma-prepaid-vs-postpaid-vaibhav-narang> (accessed on 2019-09-20)
- [15] V. B. Lestienne, "Digitization and disruption in customer relations", *Dimelo*, 2015 [Online]. Available: <https://www.dimelo.com/en/blog/digitization-and-disruption-customer-relations/> (accessed on 2019-11-11)
- [16] "Users want operators to be more like Amazon, Netflix", *Mobile World Live*, 2017 [Online]. Available: <https://www.mobileworldlive.com/featured-content/top-three/users-want-operators-to-be-more-like-amazon-netflix> (accessed on 2019-09-13)
- [17] J. Kyriakakis, "The digital enabler: a charging, self-care & marketing platform at the core of the mobile business", *The Black Swan Telecom J.*, 2016 [Online]. Available: http://bswan.org/digital_enabler.asp (accessed on 2019-02-05)
- [18] "Slow mobile apps ruin everything", *New Relic* [Online]. Available: <https://newrelic.com/resource/slow-mobile-apps-ruin-everything> (accessed on 2019-05-10)
- [19] A. A. Saunders, "How big data and analytics can drive profitability for telecom operators", *Digital Doughnut*, 2017 [Online]. Available: <https://www.digitaldoughnut.com/articles/2017/august/howbig-data-and-analytics-can-drive-profitability> (accessed on 2019-05-07)
- [20] H. Prasad, "By the numbers: three big data opportunities for telecoms", *TeleTech*, 2013 [Online]. Available: <http://www.teletech.com/resources/articles/numbers-three-big-dataopportunities-telecoms> (accessed on 2019-05-05)
- [21] "MyBusiness aplicaões on Google Play" [Online]. Available: <https://play.google.com/store/apps/details?id=com.evampsaanga.ojosefcare> (accessed on 2019-08-05)
- [22] "My Orange application" [Online]. Available: <https://www.orange.jo/en/services/pages/my-orangeapplication.aspx?utm-source=orangejoutm-medium=linkutmcampaign=homepageutm-term=homepageutm-content=homepage> (accessed on 2019-06-09)
- [23] "Mitt Telenor Google Play webpage" [Online]. Available: <https://play.google.com/store/apps/details?id=no.telenor.faktura>
- [24] "App e rea de Cliente – NOS" [Online]. Available: <http://www.nos.pt/particulares/areadecliente/Paginas/default.aspx> (accessed on 2019-06-20)
- [25] "App NOS Net – NOS" [Online]. Available: <http://www.nos.pt/particulares/internet/internet-fixa/Paginas/app-nos-net.aspx> (accessed on 2019-07-03)
- [26] "Nokia, Nokia: telecom analytics" [Online]. Available: <https://networks.nokia.com/solutions/analytics> (accessed on 2019-07-08)
- [27] "Subscription platform for telecom operators – Mahindra Comviva" [Online]. Available: <https://www.mahindracomviva.com/products/businesssolutions/centralized-subscription-manager.htm> (accessed on 2019-07-18)
- [28] "Call accounting reporting software – Calero" [Online]. Available: <https://www.calero.com/clm/call-accounting-and-reporting> (accessed on 2019-09-22)
- [29] "Telecom service assurance: integrated assurance – intelligent assurance analytics – Comarch" [Online]. Available: <https://www.comarch.com/telecommunications/serviceassurance/what-is-comarch-integrated-assurance> (accessed on 2019-09-20)
- [30] S. Aheleroff, "Customer segmentation for a mobile telecommunications company based on service usage behavior", in *Proc. 3rd Int. Conf. on Data Mining and Intell. Inf. Technol. App.*, 2011, pp. 308–313
- [31] J. Bayer, "Customer segmentation in the telecommunications industry", *J. of Database Market. & Cust. Strat. Manag.*, vol. 17, no. 3–4, pp. 247–256, 2010 (DOI: 10.1057/dbm.2010.21).
- [32] A. Arora and R. Vohra, "Segmentation of mobile customers for improving profitability using data mining techniques", *Int. J. Comput. Sc. and Inf. Technol.*, vol. 5, no. 4, pp. 5241–5244, 2014 [Online]. Available: <http://ijcsit.com/docs/Volume%205/vol5issue04/ijcsit2014050492.pdf>
- [33] NGINX webpage [Online]. Available: <https://www.nginx.com/>
- [34] Apache Tomcat webpage [Online]. Available: <http://tomcat.apache.org/>
- [35] MySQL webpage [Online]. Available: <https://www.mysql.com/> (accessed on 2019-09-15)
- [36] Redis webpage [Online]. Available: <https://redis.io/> (accessed on 2019-08-13)
- [37] "RESTful Web services", *IBM Developer* [Online]. Available: <https://developer.ibm.com/articles/ws-restful/>
- [38] JSON web token [Online]. Available: <https://jwt.io/>
- [39] R. Nagarajan and R. Thirunavukarasu, "A review on intelligent cloud broker for effective service provisioning in cloud" in *Proc. of the 2nd Int. Conf. on Intell. Comput. and Control Syst.*, 2018 (DOI: 10.1109/ICCONS.2018.8662953).
- [40] D. K. Barry, "Service oriented architecture (SOA)", *Barry Associates* [Online]. Available: <https://www.servicearchitecture.com/articles/web-services/service-oriented-architecturesoa-definition.html> (accessed on 2019-09-10)
- [41] N. Niknejad, W. Ismail, I. Ghani, B. Nazari, M. Bahari, and A. Hussin, "Understanding service-oriented architecture (SOA): a systematic literature review and directions for further investigation", *Inf. Systems*, vol. 91, 2020 (DOI: 10.1016/j.is.2020.101491).
- [42] R. Nagarajan and R. Thirunavukarasu, "Service-oriented broker for effective provisioning of cloud services – a survey", *Int. J. of Comput. and Dig. Syst.*, vol. 9, no. 5, pp. 863–879, 2020 (DOI: 10.12785/ijcds/090508).
- [43] "SOAP version 1.2 part 1: messaging framework (second edition)" [Online]. Available: <https://www.w3.org/TR/soap12/> (accessed on 2019-09-14)
- [44] "Introducing JSON" [Online]. Available: <https://www.json.org> (accessed on 2019-04-02)
- [45] J. Freeman, "What is an API? Application programming interfaces explained", *InfoWorld*, 2019 [Online]. Available: <https://www.infoworld.com/article/3269878/what-is-an-api-application-programming-interfaces-explained.html> (accessed on 2019-09-10)
- [46] "Service-oriented architecture (SOA)", *IBM* [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSMQ79_9.5.1/com.ibm.epl.pg.doc/topics/pegl_serv_overview.html (accessed on 2019-09-10)
- [47] N. Rajganes and T. Ramkumar, "A review on broker based cloud service model", *J. of Comput. and Inf. Technol.*, vol. 24, no. 3, pp. 283–292, 2016 (DOI: 10.20532/cit.2016.1002778).
- [48] T. Ramkumar, S. Hariharan, and S. Selvamuthukumaran, "A survey on mining multiple data sources", *Wiley Interdiscip. Rev. Data Min. and Knowl. Discov.*, vol. 3, no. 1, pp. 1–11, 2013 (DOI: 10.1002/widm.1077).
- [49] T. Venkatesan, K. Saravanan, and T. Ramkumar, "A big data recommendation engine framework based on local pattern analytics strategy for mining multi-sourced big data", *J. of Inform. & Knowl. Manag.*, vol. 18, no. 1, 2019 (DOI: 10.1142/S0219649219500096).
- [50] "Apache Spark webpage" [Online]. Available: <https://spark.apache.org> (accessed on 2019-04-27)
- [51] S. Anquetil, "What are push notifications and why are they important to your app?", *Business 2 Community*, 2014 [Online]. Available: <https://www.business2community.com/mobile-apps/push-notifications-important-app-01008413>
- [52] "Firebase cloud messaging", *Firebase Google* [Online]. Available: <https://www.firebase.google.com/docs/cloud-messaging/> (accessed on 2019-07-15)



Patrick Santos is working as a backend developer at TIMWE SPIN Lab. He has received his M.Sc. degree in Computer Science from the Technology High School, Castelo Branco, Portugal. He has also contributed with some scientific publications on multiple subjects, such as context-aware systems, intelligent personal assistants, and

payment ecosystems. His main research interests include ubiquitous computing and Big Data.

E-mail: patrick.santos@timwetechnology.com

TIMWE Spin Lab

Parque da Ciencia e da Tecnologia da Universidade do Porto

Porto, Portugal



Eduardo Seabra is working at TIMWE SPIN Lab as a backend developer. He received his M.Sc. in Electronic Engineering and Telecommunications from the University of Aveiro, Portugal. He was a researcher at the Instituto de Telecomunicações, with his main research interests focusing on Web 2.0 and IPTV services.

E-mail: eduardo.seabra@timwetechnology.com

TIMWE Spin Lab

Parque da Ciencia e da Tecnologia da Universidade do Porto

Porto, Portugal



Jorge Ramos is currently working as a Project Manager at TIMWE SPIN Lab, a company in Porto, Portugal. He has a B.Sc. degree in Electrical Engineering and Computer Science from Nova University in Lisbon. Jorge has considerable experience in project management in the telecommunications sector, thanks to more than

15 years spent working with telco operators. His research interests include network virtualization techniques and delay tolerant networks.

E-mail: jorge.ramos@timwetechnology.com

TIMWE Spin Lab

Parque da Ciencia e da Tecnologia da Universidade do Porto

Porto, Portugal



José Castro is currently working at TIMWE SPIN Lab as a frontend developer. He has an M.Sc. in Information Systems Engineering and Management from the University of Minho, Portugal. José boasts considerable experience in developing mobile solutions (Android and iOS) and also in developing web platforms for

customers from different sectors. His main research interests include distributed systems and computer security.

E-mail: jose.castro@timwetechnology.com

TIMWE Spin Lab

Parque da Ciencia e da Tecnologia da Universidade do Porto

Porto, Portugal

Autonomous Instrumentation for Measuring Spontaneous Electromagnetic Emissions in Mining

Krzysztof Maniak¹ and Remigiusz Mydlikowski²

¹ National Institute of Telecommunications, Warsaw, Poland

² Faculty of Electronics, Wrocław University of Science and Technology, Wrocław, Poland

<https://doi.org/10.26636/jtit.2020.145320>

Abstract—This article presents the design of an innovative receiver capable of identifying electric and magnetic components of electromagnetic fields. The receiver senses and records electromagnetic disturbances generated as mine tunnels collapse. It offers excellent operating specification and the ability to sense and log magnetic and electrical component strength values in real time. The paper analyzes the data obtained with the use of a system installed in a working mine and attempts to determine hazards resulting from increased rock stress levels that cause spontaneous EM emissions.

Keywords—coal mining, electromagnetic field, EM receiver, rock destruction.

1. Introduction

Tunnel collapses continue to be the main hazard faced in hard coal mining, posing the most serious threat to the life and health of miners [1]. Nowadays, miners are less threatened by methane explosions, due to the common use of sensitive gas detectors allowing the personnel to evacuate before critical methane concentration are reached. In the event of the collapse, weak seismic events occurring at hard coal excavation sites may be considered their potential precursors. They result from a cascade-like formation of small and large cracks that eventually lead to the collapse of rock masses. The traditional method of monitoring mine events that may potentially lead to huge rock collapses, is known as the seismic method [2]. In this method, an extensive network of geophones is used. It is operated by personnel responsible for safety of the mine, making the decision on the potential evacuation of miners.

An alternative method capable of forecasting hazards at such locations in a more effective manner consists in sensing electromagnetic field variations. It is based on the fact that rapid increase in the mechanical force exerted within rock leads to weak, spontaneous EM emissions that may be considered a precursor of a potential disaster. This phenomenon occurs also during landslides, providing positive results when correlated with the inclinometric method commonly used in landslide testing [3].

This paper aims to present the design of a compact EM field receiver capable of sensing electric and magnetic field changes, enclosed in a small housing. The receiver may also be used to expand the scope of research on the nature of EM fields generated as the mine. A computer-based analysis of the waveforms recorded that correspond to changes in field strength makes it possible to distinguish between useful signals originating from deposits subject to mechanical stress and interference signals.

Many researchers reported a rapid increase in EM field intensity during hard coal mine collapses [4]–[6]. It is important that such a phenomenon occurs even a few hours before the actual incident, i.e. it offers sufficient time for successful evacuation of the miners from the risk area. These findings were proved in laboratory tests concerned with EM field emissions originating from rock samples subjected to destructive axial pressure [7], [8]. It was observed that the electric waveforms obtained by the sensors during the tests had the form of a series of fading pulses with a continuous frequency spectrum [9]. The literature provides results of works aimed at analyzing the number of pulses within a time period, e.g. within one second, or provides descriptions of research concerned with DC voltage corresponding to instantaneous pulse values. This makes it possible to determine the amplitude of the relevant electromagnetic field components. The frequency band of EM emissions may be determined for individual rock samples tested using the method from paper [10].

2. EM Field Measurement Test Stand

In order to measure electromagnetic signals originating from rocks subjected to destructive mechanical stresses, a laboratory testbed was built, as shown in Fig. 1. Many tests of spontaneous electromagnetic emissions were performed on various types of rocks, such as coal, sandstone, dolomite and magnesite. The amplitudes of EM emissions and their spectral distributions were determined during these studies and were recorded for further analysis.

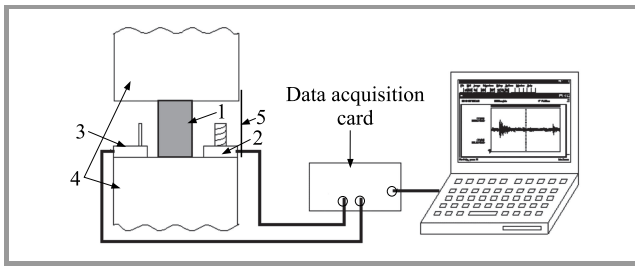


Fig. 1. Setup for recording EM fields generated by rock samples subjected to destructive loads: 1 – test sample, 2 – magnetic field sensor, 3 – electric field sensor, 4 – hydraulic press, 5 – electro-magnetic screen.

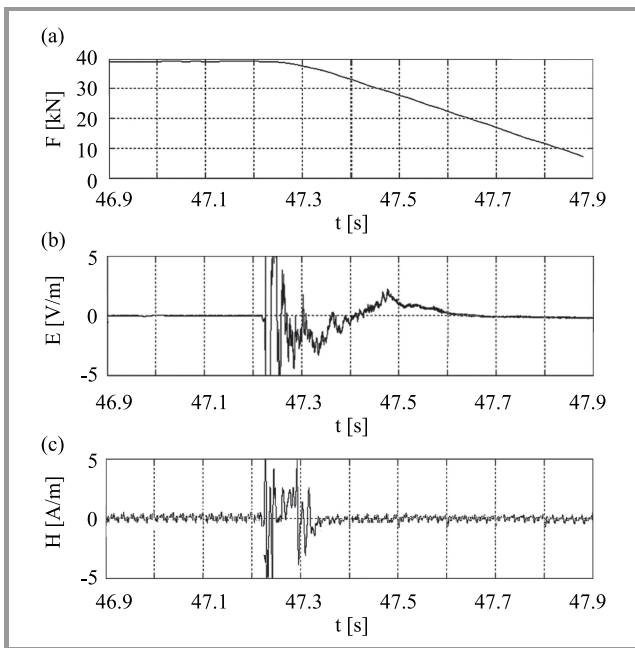


Fig. 2. The recorded EM emission from coal sample during the break process: (a) load applied, (b) electric field value, and (c) magnetic field value.

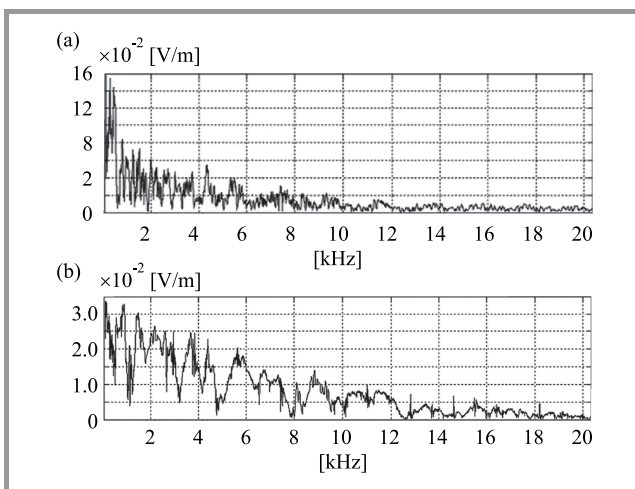


Fig. 3. Frequency spectrum of electrical (a) and magnetic (b) component of the electromagnetic wave emitted by the sample.

During the tests, rocks were placed in a hydraulic press with the maximum force of 450 kN. Two separate sensors (field intensity receivers) measuring the electric and magnetic field components were placed near the press jaws. As suggested in the literature, the broadband EM emissions recording method (up to 100 kHz) was used. To avoid interference, both measurement receivers, as well as the rock sample, were shielded using a flat, grounded plate.

Figure 2 shows examples of EM field signals recorded while testing a cylindrical hard coal sample measuring 30×50 mm. Spectral analysis of the signals is presented in Fig. 3.

3. Autonomous Receiver for EM Field Components Generated by Mine Rocks

Tests carried out in the laboratory indicated that the integrated EM field receiver and recorder, designed to operate in a coal mine environment, should comply with the following specifications [10]:

- minimum electric field component sensitivity of $E_{min} = 2 \cdot 10^{-4}$ V/m and minimum magnetic component sensitivity of $H_{min} = 1.5 \cdot 10^{-5}$ A/m,
- ability to record in the 50 kHz frequency band,
- ability to record real-time signals obtained from sensors, along with peak values,
- built-in power supply for continuous, prolonged operation, e.g. for one month,
- data protection in the event of a power failure,
- no interference with other electronic devices installed in the mine,
- immunity to electromagnetic interference occurring in the mine during its operation,
- IP67 housing and antistatic design,
- mobile enclosure that is easy to install in the mine environment.

A block diagram of the EM field receiver is shown in Fig. 4. The device is made up of four functional components:

- EM field electric signal sensor and signal conditioning circuit,
- EM field magnetic signal sensor and signal conditioning circuit,
- digital signal and control signal processing circuit,
- power supply system.

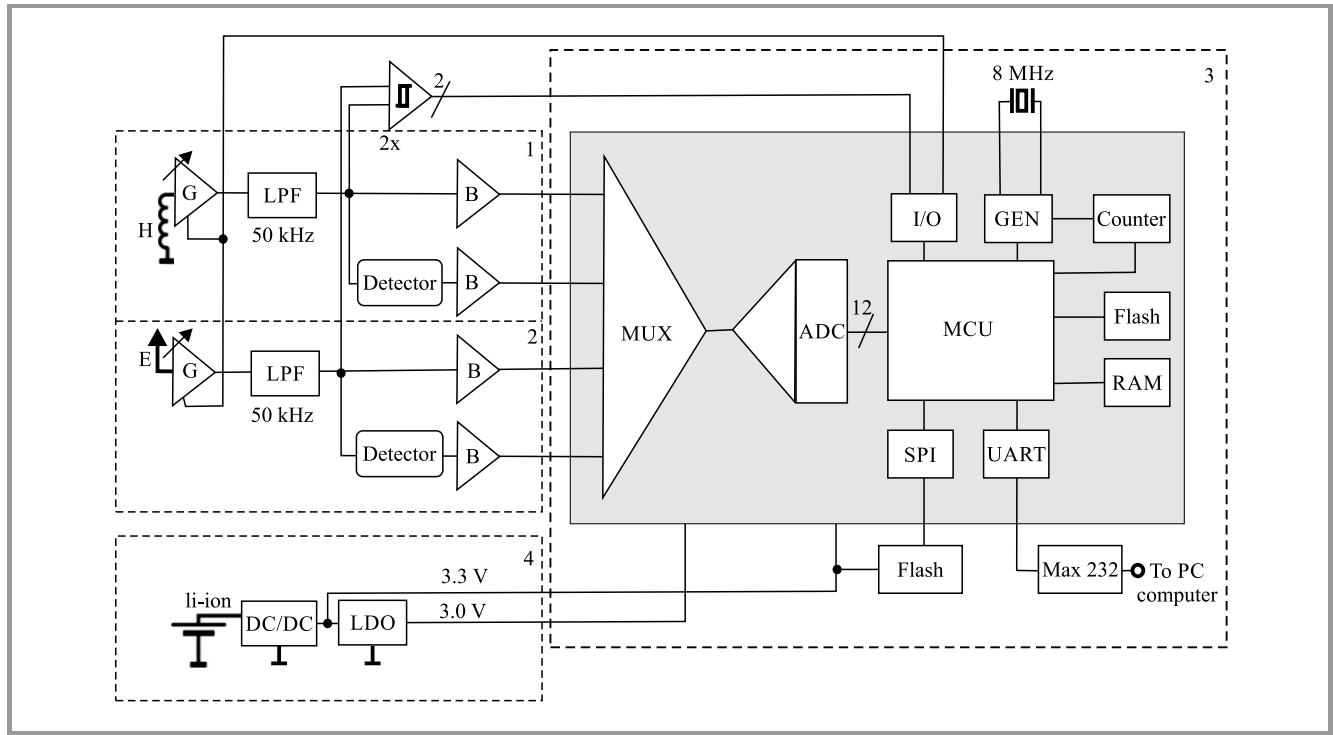


Fig. 4. Block diagram of an integrated EM receiver.

As detectors, two antennas were used: one for the magnetic component and the other for the electric component.

The effective height h_{ef} of the antenna rod is 5 cm, and because $h_{ef} \ll \lambda$ the voltage induced in the antenna is [11]:

$$SEM_E = E \cdot h_{ef} , \quad (1)$$

where E is electric field strength in [V/m].

A ferrite rod antenna coupled with an inductive coil was used as the magnetic component sensor. The induced voltage may be determined from [12], [13]:

$$SEM_H = Z_0 \cdot H \cdot \frac{2\pi \cdot z \cdot S \cdot \mu}{\lambda} , \quad (2)$$

where: H – the magnetic field strength in [A/m], $Z_0 = 120 \pi \Omega$ – the wave impedance of free space, z – the number of antenna wire turns, S – the cross section area of antenna coil in [m²], λ – the signal wavelength in [m], μ – the magnetic permeability effective value of the antenna's ferrite core in [H/m].

Sensors with the antennas are located in block 1 and 2, respectively, and form analog signal tracks. In both cases the first part of the conditioning unit has the form of a variable gain amplifier G . Next, the signals received via both tracks are fed to low-pass filters with the cut-off frequency of 50 kHz, and are then passed via buffer B to the ADC converter in MCU. The sampled signal is used to measure the instantaneous value, while the peak value detector gives two additional measurements to ADC via a MUX multiplexer. After digitalization the signals are stored on a Flash

memory card. The recorded data is directly accessible via the RS232 interface.

The digital block with a microcontroller is responsible for signal interpretation and for storing it on a Flash memory card. The MCU used is characterized by low power consumption and allows to update firmware without disassembling the device.

All electronic systems are powered from a 3.3 – 3 V power supply. Therefore, a 3.6 V lithium-ion battery was used. As a DC-DC converter and an LDO regulator are used, the analog part of the system is powered by the voltage of 3 V, while the digital part uses 3.3 V.

The three separated surface-mounted printed circuit boards are combined into a single module by relying on board-to-board mezzanine micro-connectors. The PCB with MCU acts as a base plate to which the analogue blocks with the antennas are attached.

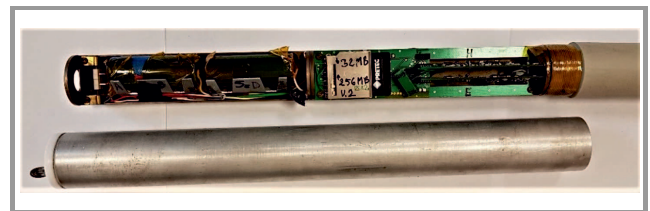


Fig. 5. The integrated electromagnetic field receiver with visible battery and PCBs.

Figure 5 presents the complete receiver unit. It is 500 mm long with 40 mm diameter.

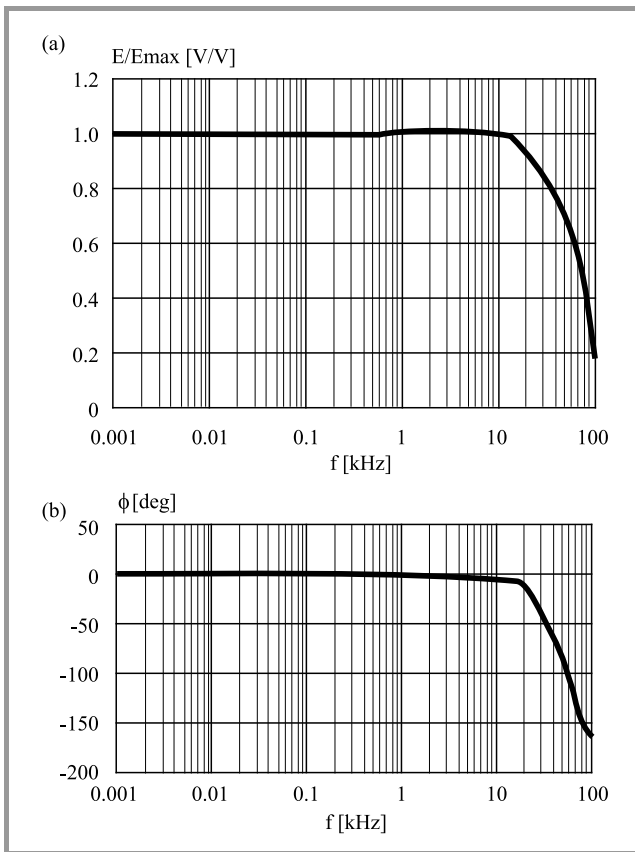


Fig. 6. Frequency characteristics of electric field sensor: (a) relative amplitude sensitivity E/E_{max} , (b) phase response.

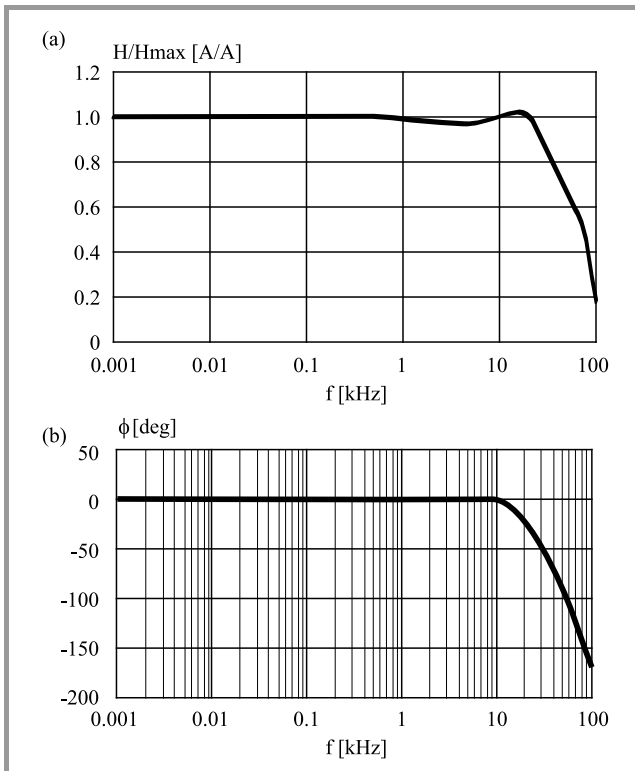


Fig. 7. Frequency characteristics of magnetic field receiver: (a) relative amplitude sensitivity H/H_{max} , (b) phase response.

3.1. Instrument Validation

A series of tests and verification measurements has been performed to determine the operating frequency range of the device's sensors. Figures 6a and 7a show frequency sensitivity of both sensing elements and their phase response. For the electric field, the upper frequency of 47 kHz was obtained, and for the magnetic receiver, it equaled 48 kHz. The sensitivity regularities observed in the operating band are below 1 dB. Out of band amplitude characteristics show the slope of -12 dB/octave, which corresponds to the double-pole transmittance type.

Figures 6b and 7b show phase characteristics of the sensing system. Both exceed the 3 dB frequency limit at the phase of -180° , which is a typical response for a double pole transmittance system [14].

3.2. Tests under Real Mining Conditions

The designed unit was installed in a real, operable hard coal mine. The receiver was placed in a borehole made in an inactive wall, where it was taking autonomous measurements over a period of 10 days.

To achieve reliable results, first the natural EM background disturbances present at the location in question were de-

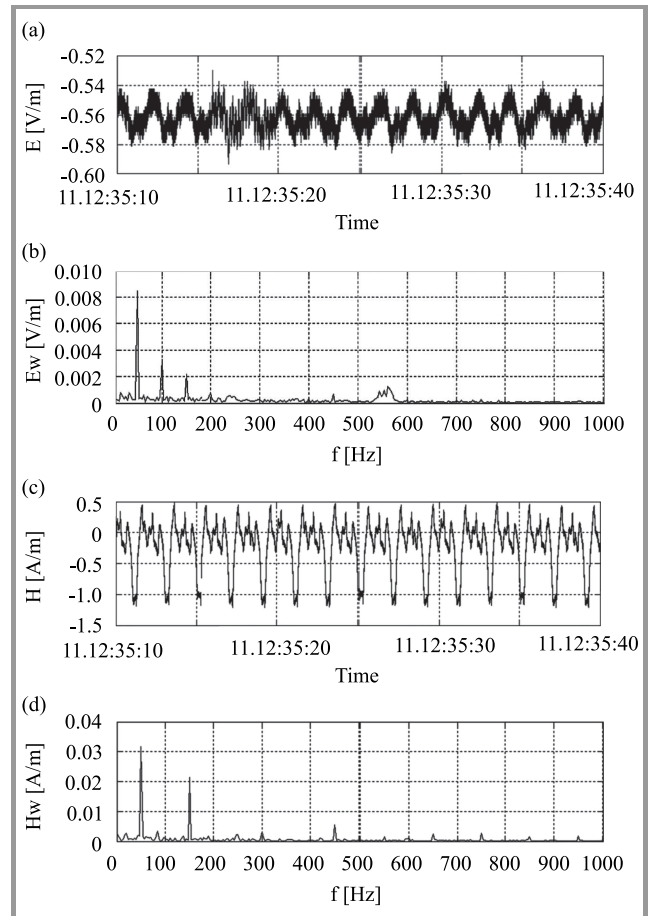


Fig. 8. EM background in an inactive mine for: (a) electric component, (b) E component spectrum, (c) H component, (d) H component spectrum.

terminated. The time waveforms obtained are periodical in nature and their peak-to-peak values are relatively low (Fig. 8a–c).

Spectral analysis shows that the majority of interferences originate from the power mains (with spectrum peaks visible at 50, 100 and 150 Hz).

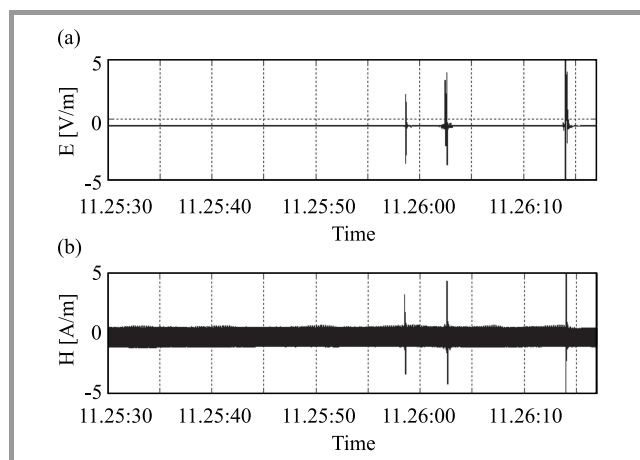


Fig. 9. Received series of EM pulses exceeding the background level registered at the measurement location.

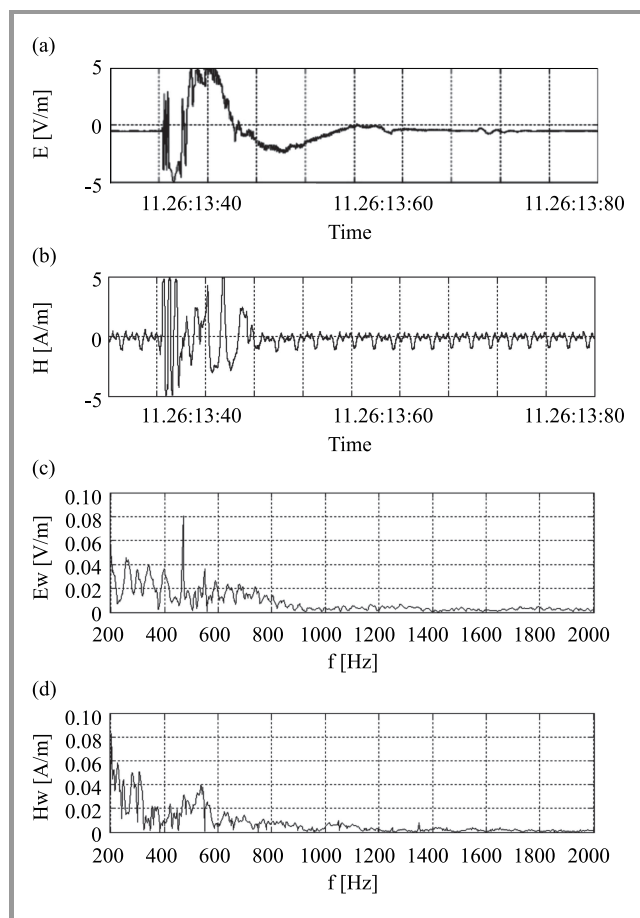


Fig. 10. Detailed analysis of recorded signal presented in Fig. 9: (a) E component, (b) H component, (c) E signal frequency spectrum, (d) H signal spectra.

During the 10-day test conducted under real mine conditions, at least several events of increased spontaneous electromagnetic emissions were recorded. They had the form of a series of pulses with different amplitudes and short repetition times. Since the receiver was installed in an inactive mine, the disturbances from equipment working at the mine could be excluded. Figure 9 shows an example of events related to a low-energy mine collapse. The results shown are representative of the entire set of EM events recorded during the test and of other seismic events that occurred in the mine.

Figure 10 presents an analysis of the series of recorded EM pulses from Fig. 9. The fading waveform is visible there (Fig. 10a–b) lasting up to 200 ms and up to 100 ms for the E component and for the H component, respectively. After performing the Fourier transformation (Fig. 10c–d), it was determined that a significant portion of the frequency response is within the 1 kHz band.

4. Conclusions

The use of an integrated, autonomous EM receiver in a real mine allows for early detection of upcoming seismic events. During its operation in the mine, the receiver recorded several such events. A single event of this type is characterized by a series of pulses with short repetition intervals. The period of the recorded pulses depends on the dynamics of the seismic process. Low-energy seismic events were recorded during mining. Interference from the power mains adversely affected the quality of the recorded data. However, with proper shielding of the receiver's sensing device, it is possible to significantly reduce the interference level. In addition, the digitized signal obtained should be processed to lower the impact of noise.

The frequency band of the recorded seismic pulses increases up to several kilohertz. This was also confirmed during laboratory tests performed as part of the research project.

Measurements of EM emissions generated by force-stressed rocks may be used as an additional warning factor. Currently, the seismic method is relied upon in mines for this particular purpose. However, new detection methods are being developed constantly and may offer a valuable addition to the classic approach. Fortunately, it is possible to create an early warning system for identifying mine collapse hazards by using a network of EM field receivers distributed throughout the mine and installed at the locations that are most sensitive, from the geological point of view.

References

- [1] M. Akgün, "Coal mine accidents", *Turkish Thoracic J.*, vol. 16, no. S1–S2, 2015 (DOI: 10.5152/ttd.2015.008).
- [2] J. L. X. Meng, Y. Wang, and Z. Yang, "Prediction of coal seam details and mining safety using multicomponent seismic data: A case history from China", *Geophysics*, vol. 81, no. 5, pp. B149–B165, 2016 (DOI: 10.1190/geo2016-0009.1).

Information for Authors

Journal of Telecommunications and Information Technology (JTIT) is published quarterly. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review.** Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

Suitable research-related papers should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published only by invitation.

Manuscript. TEX and LATEX are preferable, standard Microsoft Word format (.doc) is acceptable. The authors JTIT LATEX style file is available:

<https://www.itl.waw.pl/en/jtit-for-authors>

Papers published should contain up to 10 printed pages in LATEX authors style (Word processor one printed page corresponds approximately to 6000 characters).

The manuscript should include an abstract about 150–200 words long and the relevant keywords. The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.

Keywords should not repeat the title of the manuscript. About four keywords or phrases in alphabetical order should be used, separated by commas.

The original files accompanied with pdf file should be submitted by e-mail: redakcja@itl.waw.pl

Figures, tables and photographs. Original figures should be submitted. Drawings in Corel Draw and PostScript formats are preferred. Figure captions should be placed below the figures and can not be included as a part of the figure. Each figure should be submitted as a separated graphic file, in .cdr, .eps, .ps, .png or .tif format. Tables and figures should be numbered consecutively with Arabic numerals.

Each photograph with minimum 300 dpi resolution should be delivered in electronic formats (TIFF, JPG or PNG) as a separated file.

References. All references should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. Samples of correct formats for various types of references are presented below:

- [1] Y. Namihiro, Relationship between nonlinear effective area and mode field diameter for dispersion shifted fibres, *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.
- [2] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.
- [3] S. Demri and E. Orłowska, Informational representability: Abstract models versus concrete models, in *Fuzzy Sets, Logics and Knowledge-Based Reasoning*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314

Biographies and photographs of authors. A brief professional authors biography of up to 200 words and a photo of each author should be included with the manuscript.

Galley proofs. Authors should return proofs as a list of corrections as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within one week after receiving the offprint.

Copyright. Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript, the author(s) agree to automatically transfer the copyright for their article to the publisher, if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and all translation rights. No part of the present JTIT should not be reproduced in any form nor transmitted or translated into a machine language without prior written consent of the publisher.

For copyright form see:

<https://www.itl.waw.pl/en/jtit-for-authors>

A copy of the JTIT is provided to each author of paper published.

(Contents Continued from Front Cover)

**An Online Stream Monitoring Algorithm for Fraud Detection
in the Transport of Goods**

P. M. Białoń

Paper

79

**A Novel System Architecture for an Improved Self-care
Solution – Conceptual Design and Key Components**

P. Santos, J. Ramos, E. Seabra, and J. Castro

Paper

88

**Autonomous Instrumentation for Measuring Spontaneous
Electromagnetic Emissions in Mining**

K. Maniak and R. Mydlikowski

Paper

98



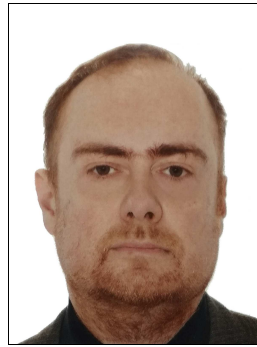
National Institute
of Telecommunications

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

tel. +48 22 512 81 83
fax: +48 22 512 84 00
e-mail: redakcja@itl.waw.pl
<http://www.nit.eu>

- [3] A. Pralat, K. Maniak, and I. Pompura, "Electromagnetic phenomena in landslides", *Acta Geodynamica and Geomaterialia*, vol. 2, no. 3, pp. 131–138, 2005 [Online]. Available: https://www.irms.cas.cz/materialy/acta_content/2005_03/16_Pralat.pdf
- [4] V. Frid, "Calculation of electromagnetic radiation criterion of rock-burst hazard forecast in coal mines", *Pure and Applied Geophysics*, vol. 158, pp. 931–944, 2001 (DOI: 10.1007/PL00001214).
- [5] V. Frid and K. Vozoff, "Electromagnetic radiation induced by mining rock failure", *Int. J. of Coal Geology*, vol. 64, no. 1–2, pp. 57–65, 2005 (DOI: 10.1016/j.coal.2005.03.005).
- [6] D. Lin-ming, L. Cai-ping, M. Zong-long, and G. Ming-shi, "Prevention and forecasting of rock burst hazards in coal mines", *Mining Sci. and Technol. (China)*, vol. 19, no. 5, pp. 585–591, 2009 (DOI: 10.1016/S1674-5264(09)60109-5).
- [7] L. Guang-Jian, Lu Cai-Ping, W. Hong-Yu, L. Peng-Fei, and Liu Yang, "Warning method of coal bursting failure danger by electromagnetic radiation", *Shock and Vibration*, vol. 2015, 2015 (DOI: 10.1155/2015/583862).
- [8] E. Wang, H. Jia, D. Song, N. Li, and W. Qian, "Use of ultra-low-frequency electromagnetic emission to monitor stress and failure in coal mines", *Int. J. of Rock Mechanics & Mining Sciences*, vol. 70, pp. 16–25, 2014 (DOI: 10.1016/j.ijrmms.2014.02.004).
- [9] Z. Shijie, S. Xiaoyu, L. Chengwu, X. Xiaoxuan, and X. Zhuang, "The analysis of coal or rock electromagnetic radiation (EMR) signals based on Hilbert-Huang transform (HHT)", in *Procedia Engin.*, vol. 26, pp. 689–698, 2011 (DOI: 10.1016/j.proeng.2011.11.2224).
- [10] R. Mydlikowski and K. Maniak, "Measurement of electromagnetic field component emission as a precursor of emerging hazard in coal mines", *J. of Telecommun. and Inf. Technol.*, no. 4, pp. 30–35, 2019 (DOI: 10.26636/jtit.2019.135419).
- [11] Z. Loni, H. Espinosa, and D. Thiel, "Insulated wire fed floating monopole antenna for coastal monitoring", *Radioengineering*, vol. 27, no. 1, pp. 127–133, 2018 (DOI: 10.13164/re.2018.0127).
- [12] V. Dyo, T. Ajmal, B. Allen, D. Jazani, and I. Ivanov, "Design of a ferrite rod antenna for harvesting energy from medium wave broadcast signals", *The J. of Engin.*, vol. 2013, no. 12, pp. 89–96, 2013 (DOI: 10.1049/joe.2013.0126).
- [13] T. Bolton and M. B. Cohen, "Optimal design of electrically-small loop receiving antenna", *Progress In Electromagnetics Research C*, vol. 98, pp. 155–169, 2020 (DOI: 10.2528/PIERC19090911).
- [14] U. Tietze, Ch. Schenk, and E. Gamm, *Electronic Circuits—Handbook for Design and Application 2nd Edition*. Springer, 2011 (ISBN-13: 9783540004295).



Krzysztof Maniak received his M.Sc. and Ph.D. degrees in Electronics from Wrocław University of Technology, Poland, in 2001 and 2008, respectively. Since 2012, he has been with the National Institute of Telecommunications in Wrocław. He is an author and co-author of numerous research papers and publications. His professional work focuses on: power systems, electromagnetic compatibility and design of measuring equipment. He also participates in research and measurement-related activities. His research interests focus on the testing of power grid quality and its broadband interference protection.

 <https://orcid.org/0000-0003-1974-909X>

E-mail: k.maniak@itl.waw.pl

National Institute of Telecommunications

Swojczycka 38

51-501 Wrocław, Poland



Remigiusz Mydlikowski received his M.Sc. and Ph.D. degrees in Electrical Engineering from Wrocław University of Science and Technology, Wrocław, Poland, in 1997 and 2002, respectively. He is currently with the Department of Theory of Field Theory, Electronic Systems and Optoelectronics at Wrocław University of Science and Technology. His research interests are in non-invasive studies of ground and rock properties.

 <https://orcid.org/0000-0002-8670-2913>

E-mail: remigiusz.mydlikowski@pwr.edu.pl

Wrocław University of Science and Technology

Wybrzeże Wyspiańskiego 27

50-370 Wrocław, Poland