

# Research in Computing Science

ISSN: 1870-4069

Vol. 143  
October 2017

# Research in Computing Science

---

## Series Editorial Board

### Editors-in-Chief:

*Grigori Sidorov, CIC-IPN, Mexico*  
*Gerhard X. Ritter, University of Florida, USA*  
*Jean Serra, Ecole des Mines de Paris, France*  
*Ulises Cortés, UPC, Barcelona, Spain*

### Associate Editors:

*Jesús Angulo, Ecole des Mines de Paris, France*  
*Jihad El-Sana, Ben-Gurion Univ. of the Negev, Israel*  
*Alexander Gelbukh, CIC-IPN, Mexico*  
*Ioannis Kakadiaris, University of Houston, USA*  
*Petros Maragos, Nat. Tech. Univ. of Athens, Greece*  
*Julian Padget, University of Bath, UK*  
*Mateo Valero, UPC, Barcelona, Spain*  
*Olga Kolesnikova, ESCOM-IPN, Mexico*  
*Rafael Guzmán, Univ. of Guanajuato, Mexico*  
*Juan Manuel Torres Moreno, U. of Avignon, France*

### Editorial Coordination:

*Alejandra Ramos Porras*

*Research in Computing Science*, Año 16, Volumen 143, octubre de 2017, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. <https://www.rcs.cic.ipn.mx>. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de octubre de 2017.

Las opiniones expresadas por los autores no necesariamente reflejan la postura del editor de la publicación.

Queda estrictamente prohibida la reproducción total o parcial de los contenidos e imágenes de la publicación sin previa autorización del Instituto Politécnico Nacional.

*Research in Computing Science*, year 16, Volume 143, October 2017, is published monthly by the Center for Computing Research of IPN.

The opinions expressed by the authors does not necessarily reflect the editor's posture.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research of the IPN.

# Computing & Telematics

**Roberto Zagal**  
**Félix Mata**  
**Carlos Hernández (eds.)**



Instituto Politécnico Nacional  
“La Técnica al Servicio de la Patria”



Instituto Politécnico Nacional, Centro de Investigación en Computación  
México 2017

**ISSN: 1870-4069**

---

Copyright © Instituto Politécnico Nacional 2017  
Formerly ISSN: 1665-9899

Instituto Politécnico Nacional (IPN)  
Centro de Investigación en Computación (CIC)  
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal  
Unidad Profesional “Adolfo López Mateos”, Zacatenco  
07738, México D.F., México

<http://www.rcs.cic.ipn.mx>  
<http://www.ipn.mx>  
<http://www.cic.ipn.mx>

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

All rights reserved. No part of this publication may be reproduced, stored on a retrieval system or transmitted, in any form or by any means, including electronic, mechanical, photocopying, recording, or otherwise, without prior permission of the Instituto Politécnico Nacional, except for personal or classroom use provided that copies bear the full citation notice provided on the first page of each paper.

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

## Table of Contents

	Page
Testing Environment for Precision Agriculture Supported in IoT .....	7
<i>E. A. Quiroga, S. F. Jaramillo, W. Y. Campo Muñoz, G. E. Chanchí Golondrino, C. L. Burbano G.</i>	
Design and Implementation of a Stand-Alone Home Energy Management System Based on Internet of Things .....	17
<i>Javier E. Sierra, Boris Medina, Ramón Álvarez</i>	
Analysis of Feature Sets for Malware Family Classification.....	27
<i>Jesús Javier Reyes Torres, Eleazar Aguirre Anaya, Ricardo Menchaca Méndez, Nareli Cruz Cortés, David Alejandro Robles Ramírez</i>	
Ecosystems of Collaborative Learning in Educational Technological Mediations: Case Analysis .....	41
<i>Carolina Burbano Gonzalez, Clara Burbano Gonzalez, Katerine Márceles Villalba</i>	
Cryptoanalysis of the 340-bit RSA Algorithm using SBC .....	45
<i>Nelson Darío Pantoja, Anderson Felipe Jiménez, Siler Amador Donado, Katerine Márceles</i>	
Security Validation with Owasp Mobile for the Data Protection in Oral Health .....	57
<i>Katerine Márceles, Clara L. Burbano, Gustavo Uribe, Diana Burbano</i>	
Conceptualization of Serious Games Used for the Dissemination of the Historical and Cultural Heritage in Colombia .....	65
<i>Claudia Sofía Idrobo, Maria Isabel Vidal, Katherine Marceles, Clara Lucía Burbano</i>	
Proposal of an Interconnection Management Model and Availability of Internet Access for Things (MGID-IoT).....	77
<i>Chadwick Carreto, Jhovanny Pasaran, Alfonso Fernandez</i>	
Architecture of Mobile Services applied to the Internet of Things .....	85
<i>Chadwick Carreto, Francisco Cerda, Felipe Menchaca</i>	
Embedded System for the Regular Blood Pressure Monitoring .....	93
<i>V. H. Garcia, N. Vega, R. Hernandez</i>	

Cryptography in Wireless Network Penetration Testing .....	105
<i>Claudio Casado, Cristian Barría, Lorena Galeazzi</i>	
Computer Vision Algorithm Implementing Geometric Morphometry to the Shape Analysis .....	113
<i>L. Jhonathan Flores-Guarneros, B. Esther Carvajal-Gómez</i>	
Application of the Shamir Threshold Scheme to a System for Safely Storing and Sharing Experimental Clinical Studies in Accordance with the Official Mexican Standard NOM-024-SSA3-2012.....	121
<i>José Daniel Pérez Ramírez, Lorena Chávez Nava Olguín, Blanca Alicia Rico Jiménez, Carlos Hernández Nava, Laura Ivoone Garay Jiménez</i>	
Proposal of a Communication Architecture for the Configuration and Monitoring of an Electric Microgrid .....	133
<i>V. H. Garcia, R. Ortega, R. Hernández, M.A. Ramírez</i>	
Proposal of Architecture for the Monitoring of Vital Signs based on Embedded Systems.....	143
<i>Jorge Martínez, Víctor García, Rubén Ortega</i>	
Obfuscated Information Classification .....	153
<i>Florencio Javier González Rodríguez, Eleazar Aguirre Anaya, Moisés Salinas Rosales, Ricardo Barrón Fernández</i>	
Mobile computing system for neuropsychological evaluation .....	167
<i>Enrique Alfonso Carmona García, Elena Fabiola Ruiz Ledesma, Laura Ivoone Garay Jiménez, Mario Eduardo Rivero Ángeles</i>	
Body Sensor Network Using a Domotic System .....	177
<i>Francisco Beltrán, Felix Mata, Mario Rivero</i>	
Price Forecasting: A Data Science Approach.....	183
<i>Paul Ricardo Millán González, Félix Mata</i>	

*Table of Contents*



# Testing Environment for Precision Agriculture Supported in IoT

E. A Quiroga<sup>1</sup>, S. F. Jaramillo<sup>1</sup>, W. Y. Campo Muñoz<sup>1</sup>, G. E. Chanchí Golondrino<sup>2</sup>,  
C. L. Burbano G.<sup>3</sup>

<sup>1</sup> Universidad del Quindío, Armenia- Quindío, Colombia

{eaquirogam, sfjaramillo}@uqvirtual.edu.co

wycampo@uniquindio.edu.co

<sup>2</sup> Institución Universitaria Colegio Mayor del Cauca, Popayán- Cauca, Colombia

gchanchi@unimayor.edu.co

<sup>3</sup> Fundación Universitaria de Popayán, Popayán- Cauca, Colombia

claritaluciab@gmail.com

**Abstract.** The technification of agriculture from the field of IoT allows to contribute to improve the agricultural productivity, from the provision of meteorological forecasts in an agile way using networks of sensors for the real time monitoring of multiple variables. However, a large number of sensors, capture devices and data processing technologies are available for the implementation of this type of system, which makes it necessary to choose the most suitable technologies for the experimentation and construction of IoT systems focused on agriculture accuracy. In this paper we present a test environment for precision agriculture, which was configured from a set of hardware and free software tools associated with IoT. The proposed environment took into account the Lambda architecture and aims to serve as a guide for the implementation of services in scenarios of precision agriculture.

**Keywords:** Climate Variables; Internet of Things; Lambda Architecture; Precision Farming; Weka.

## 1 Introduction

Crop management supported by precision agriculture tools encompasses monitoring activities, decision support tools and actions that automatically control one or more systems (irrigation, frost protection, fertilization, etc.). (GPS), wireless sensor networks, drones, multiple electronic devices, and the application of computer tools from Machine Learning (IoT Simple, 2017). Through this wide range of resources, a farmer can obtain detailed crop information, soil conditions and even more granular climatic variations than traditional farming techniques could not provide; Impacting the quality of the products, the processes that are carried out and the raw materials used in the activity (Solutek, s.f.).

This technification required in precision agriculture environments can be carried out through Internet of Things (IoT). In IoT things, that is, embedded devices can be ideally available anytime, anywhere. In a more technical sense, IoT consists of the integration of sensors and devices into everyday objects so that they are connected to the Internet through fixed and wireless networks (Fundación de la Innovación Bankinter, 2011) (Ruiz, 2016). Cost reduction, improvements in crop processing and care, optimization of the use of material and human resources, increased yield per hectare cultivated, higher quality of final product and reduction of disposal, Compliance with national and international requirements of production and product characteristics, etc., are some of the benefits of opting for a precision agriculture solution supported by IoT.

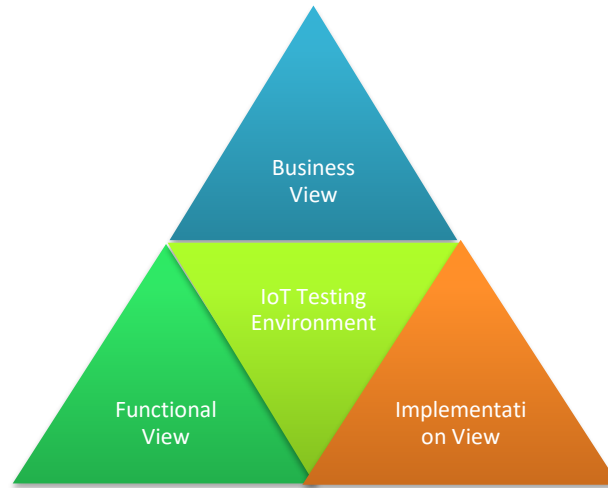
In order to achieve these objectives in the agriculture sector, nowadays a large number of sensors, capture devices and data processing technologies are available, making it necessary to choose the most suitable technologies for the experimentation and construction of IoT systems focused on the agriculture of precision. However, the problem for a farmer who has not yet implemented some of these precision agriculture tools in his crops is the high cost and complexity in interpreting the information obtained from these technologies. Due to this problem, there is a great demand and need to integrate a set of appropriate and low-cost components, given the large number of technologies that are emerging to simplify the collection, storage and processing of climatological data (Lopez, Chavez, & Sánchez, 2017) (Taffernaberry, Diedrichs, Pérez, Pecchia, & Tabacchi, 2016).

Taking into account the above problem, in this paper we present as main contribution a test environment for precision agriculture supported in open technologies for obtaining, visualization and real time analysis of climatic variables in a small scale test crop. This testing environment is addressed from three high-level views associated with the business model, functionality and implementation of the environment for precision agriculture farming scenarios. Each of the views has taken as reference the layer structure of the Lambda architecture (data capture layer, storage layer, processing layer and query layer) (Deshmane, 2015). This testing environment aims to serve as a reference for the implementation of services in the area of agriculture, in order to improve agricultural productivity, through the flexibility and functionality offered by the combination of free hardware and software tools for implementation of services supported in IoT. The rest of the article is organized as follows: section 2 describes each view of the test environment; section 3 presents the evaluation of the test environment and finally section 4 presents the conclusions and future work derived from the present research.

## **2 Test Environment Supported in IoT**

This section describes the test environment for precision agriculture, which is represented by three views (Jimenez, Hincapié, & Quintero, 2016): business view, functional view and implementation view (see Fig 1). The business view presents the business model of the test environment in the context of precision agriculture. In the functional view the different functional components of the test environment are

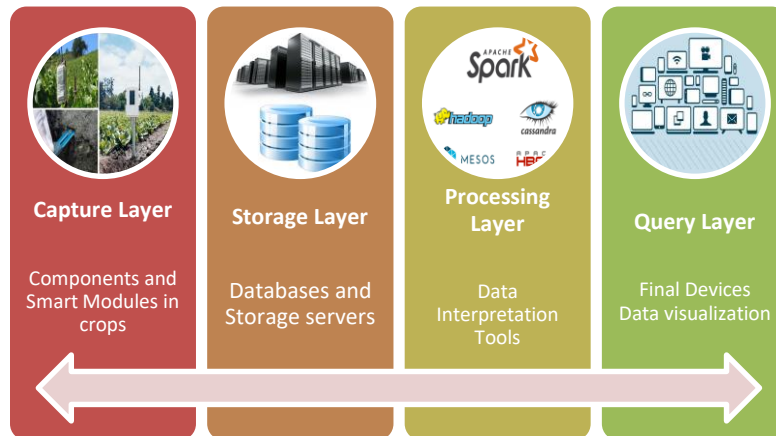
described and in the implementation view are presented the different hardware and software components that were chosen for the construction of the environment.



**Fig.1.** Diagrams of the IoT test environment.

## 2.1 Business View

This business view presents the proposed IoT testing environment from the four layers of the Lambda architecture: capture layer, storage layer, processing layer and query layer (see Fig. 2).



**Fig. 2.** Business diagram for the IoT test environment.

As shown in Fig. 2, the capture layer includes sensors and data acquisition modules, as well as embedded systems, microcontrollers or meteorological stations for the taking

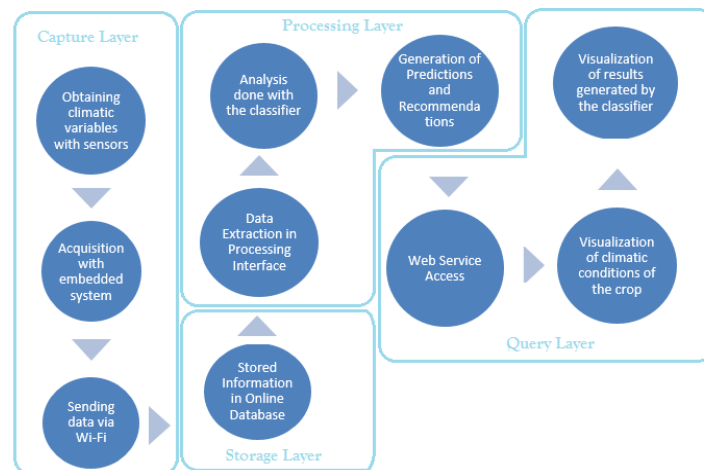
of climatic conditions in the crop. The storage layer consists of platforms, servers and database that allow to save all the data obtained in the capture layer. In the processing layer, these stored data are extracted to perform their analysis and obtain information of interest that can be applied to the crops. The final component in the environment is the query layer, which allows the farmer to visualize in real time the behavior of the variables associated with the crop, as well as the predictions and recommendations to be taken according to the needs of the crop.

## 2.2 Functional View

In Fig. 3, the different functional modules associated with the different layers of the test environment are presented. In the capture layer, the incoming data stream is taken with the data capture system (sensors-microcontroller) and can be sent in two directions, either the storage layer (online database, cloud, servers) or to the processing layer (machine learning tools). This information can be consulted through internet using final devices (web services, mobile applications) by means of the query layer.

The storage or security layer provides a history file, saving all the data that has been collected. Its storage could be forever, or partially. Storage forever would support advanced analysis and the predictive models to be implemented would have greater accuracy in creating signals on the transmission platform and for general queries. The rate or processing layer is defined as a combination of queue holdings, streaming and operational data by performing real-time analysis. The storage layer bases its estimates on previously stored data, so it is the responsibility of the speed layer to obtain real-time analysis from fast-moving data.

The functions described for each of the layers of the test environment can be broken down into functional sub-modules as presented in Fig. 3, which presents the functional sequence of the environment supported in IoT for precision agriculture, according to the layers provided by the Lambda architecture



**Fig. 3.** Functional sequence of the IoT test environment.

### 2.3 Implementation View

In this section we present in detail the hardware and software components that were chosen for the implementation of the IoT supported test environment described in this article (see Fig. 4), which are organized according to the layers of the Lambda architecture. According to Fig. 4, in "1", sensors of temperature, relative humidity, soil moisture and luminosity are placed in the test culture to realize the corresponding climatic variables. In "2" by means of the IoT Arduino YÚN board (Schwartz, 2014), these sensors are configured with their necessary parameters and wireless communication is established, thus forming the data capture layer. Through the capture board is sent in real time the registration of each of the variables associated with the sensors to be stored in "3", through the online database Ubidots (Torquica & Guzmán, 2016). In "4" using the Spark web development framework, a web service is implemented in Java that allows the online processing of the data captured from the previous layers.

To perform the processing of this data, the Spark framework integrates the API provided by the Weka data mining tool into "5". Using this API, a classifier based on the Naive Bayes algorithm is implemented with the information contained in the file in arff (attribute-relation file format) format, obtaining the Processing layer, which provides predictions of the atmospheric conditions of the crop as well as recommendations for applying inputs. Finally, in "6" the end users access the web service to consult the predictions and recommendations generated by the classifier, thus obtaining the query layer. The following is a detailed description of each layer of the implementation view.

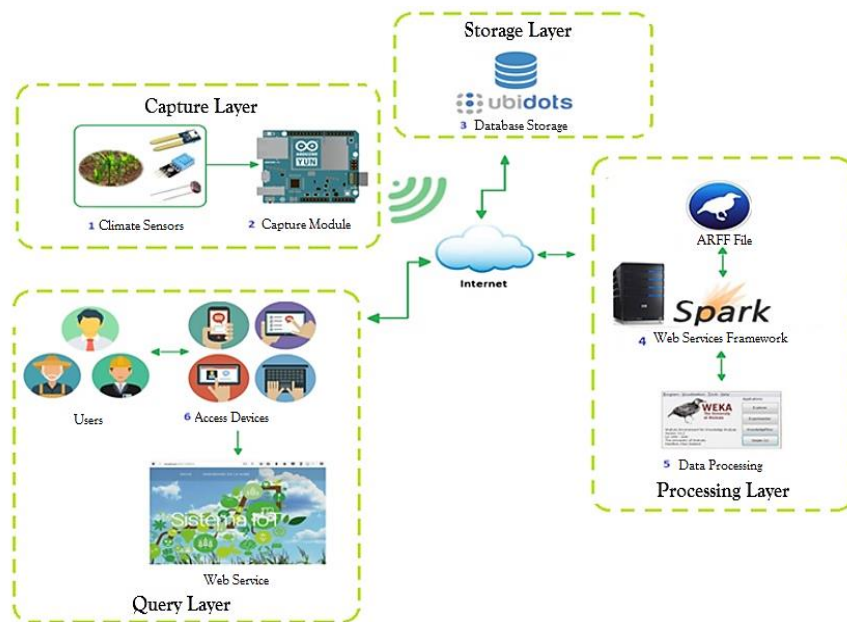


Fig. 4. Implementation diagram of the IoT environment.

### Capture Layer

Fig. 5 shows the connection scheme established to operate the sensors of temperature, relative humidity, soil moisture and luminosity connected next to the Arduino YÚN board for the capture of climatic variables.

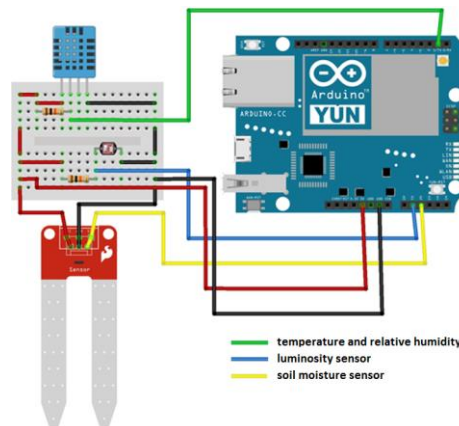


Fig. 5. Schemes of connection of sensors with microcontroller.

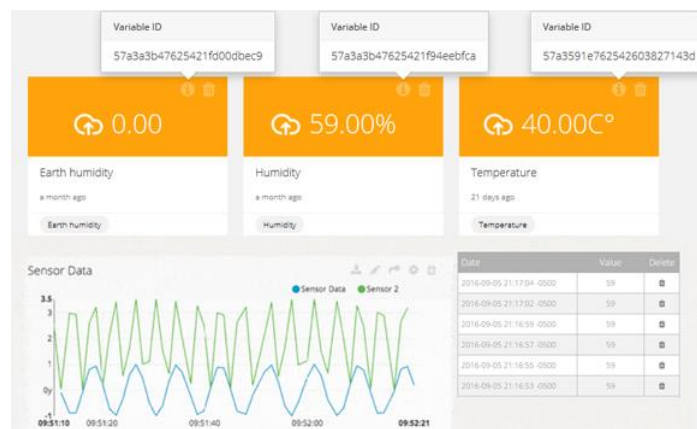


Fig. 6. Associated climatic variables in online database and their storage.

For this research the IoT Arduino YÚN board was used as the appropriate embedded system in the data capture layer and for use in the design of connected devices and in IoT projects. This board combines the power of Linux with the ease of use of Arduino since having communication and configuration via Wi-Fi. In addition it allows to send the data of climatic variables wirelessly to the other layers of the proposed environment. The sensors used were the DHT11, SEN92355P and a LDR, that were chosen for their low cost to meet the requirement of economic components for the implementation of the test environment.

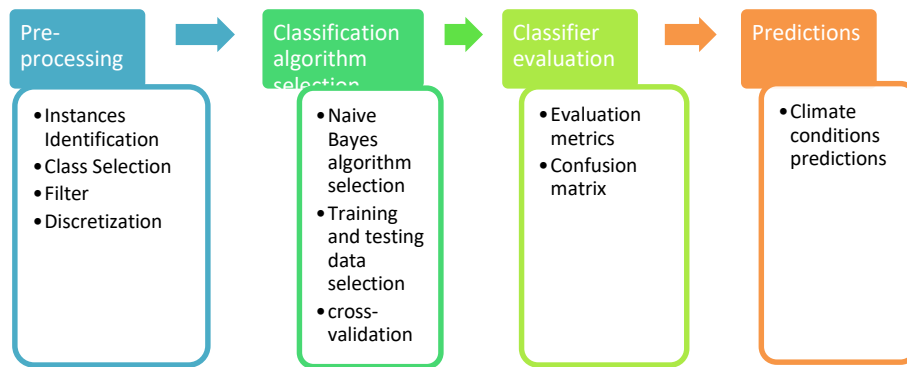
### Storage Layer

To establish the Lambda-based IoT testing environment, you must have a storage manager that serves as the storage layer. For this, we selected the online database Ubidots which is a specialized tool for IoT applications, which offers a platform for developers. Ubidots allows you to easily capture sensor data to become useful information, as well as support for the Arduino YUN capture module. Fig. 6 shows the Ubidots administration interface with the different climatic variables captured through the Arduino YÚN board.

### Processing Layer

This layer comprises two stages of processing: the generation of a file in the specific format of the Weka software and the processing of the generated file making one of the Weka data mining software:

**ARFF file generation.** To process climate variables with the Weka data mining tool, information must be organized in a specific way, this is done by an ARFF file. The list of attributes or meteorological variables to be considered (humidity, relative humidity, temperature, etc.) and the class in which the attributes of these attributes are classified are used in the header of the ARFF file, which is used to generate irrigation predictions. In the body of the document are located the different instances of these attributes, which are fed with the values obtained from the sensors of the crop.



**Fig. 7.** Weka data mining software operation process.

**Analysis with Weka software.** Once the ARFF file is generated, it is analyzed through the API provided by the Weka tool, following the sequence of steps of Fig. 7. The first stage is the pre-processing, which consists of the manipulation of vectors with discrete content of the sensed values, where loading the file with the necessary data, identifies the class to predict and can be filtered attributes, instances or discretization. The second stage consists of the generation of the classifier using a learning algorithm to which a

set of data is given which correspond to the contents of the ARFF file. For this investigation, the Naive Bayes algorithm was used and the data provided to generate the classifier were selected as a training data set and with cross validation. In the third stage the evaluation of this classifier is done, which can be defined as the degree of agreement between the values of the class assigned by the classifier and their correct locations according to the data of the ARFF file.

### Query Layer

To complete the final stage of the testing environment, a graphical GUI access interface was developed by the Spark framework. The interface developed allows real-time visualization of the climatic data of temperature, relative humidity, soil moisture and illumination being taken by the data capture component, see Fig. 8. In addition, this interface makes it possible to consult the history online with the Data stored in the Ubidots database, as well as the predictions and recommendations associated with the crop.

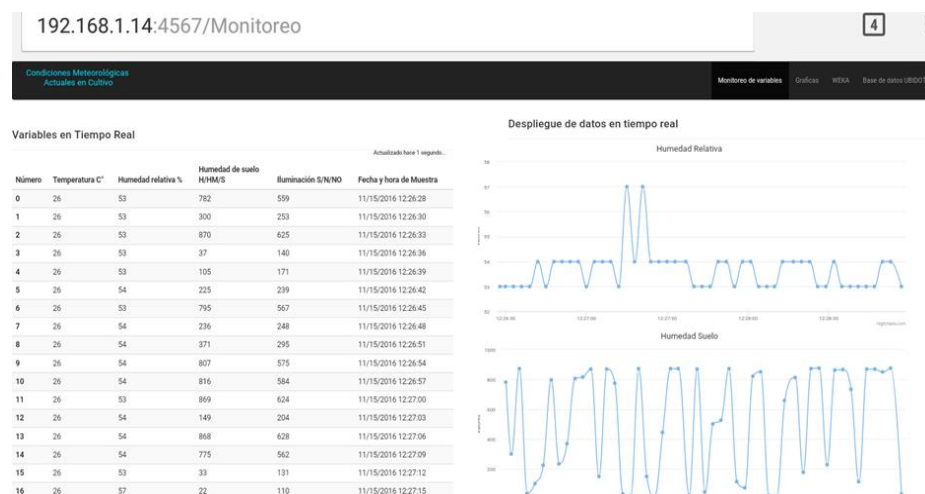
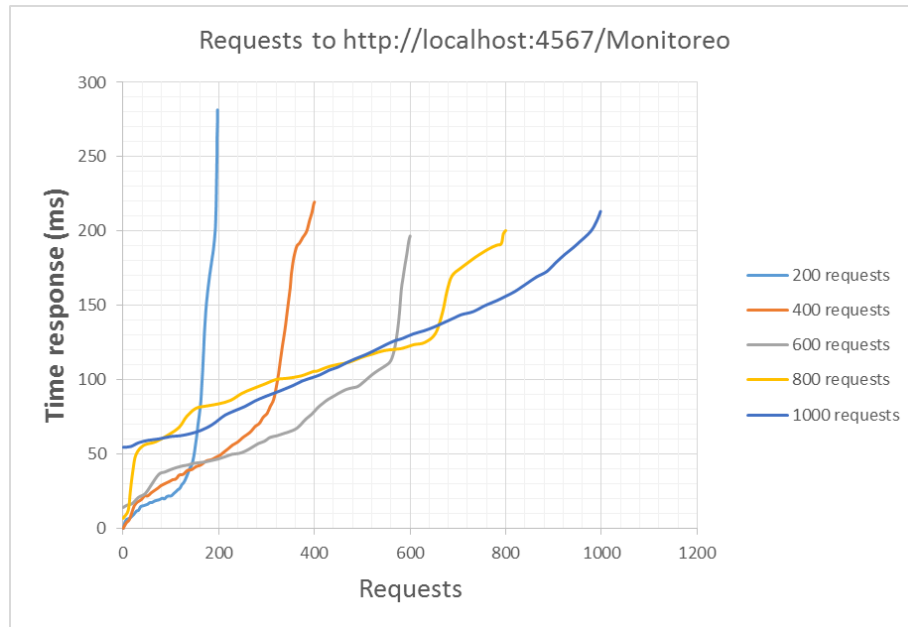


Fig. 8. Visualization of the web service for the query of the information.

## 3 Evaluation of the IoT Test Environment

Finally, by way of evaluation of the testing environment, the main component of the environment was validated, which corresponds to the processing module, which is responsible for extracting and processing information from the Ubidots database. This component was subjected to load tests in order to determine its capacity and response time. These tests make it possible to obtain the limits under which the processing module will behave appropriately and let you know when different optimization strategies need to be implemented. Thus, Fig. 9 shows the response times for the stress tests performed on the processing module, which were carried out by incremental variation from 100 to 1000 concurrent requests.



**Fig. 9.** Response times for stress test in processing layer.

Fig. 9 shows that the maximum time obtained in each of the concurrent request series (200,400,600,800,1000) varies between 200 and 260 milliseconds. In the same way, in each of the series as the number of requests increases, the time grows proportionally; being the series of 200 requests the greatest slope, and the series of 1000 petitions the smallest one, which allows to conclude that the processing module presents an appropriate behavior as the number of concurrent requests grow. Thus, for the particular case of 1000 concurrent request series, the response time increases from 160 ms to 220 ms between 800 and 1000 requests, which represents an average of 0.3 ms for each additional concurrent request.

## 4 Conclusions and Future Work

Taking advantage of the great flexibility and functionality offered by the combination of low cost hardware tools and open source software, we have achieved the design, implementation and deployment of an IoT supported test environment, which is addressed from three design views. In addition, four layers of operation are deployed, including the capture of climatic variables, the storage of this information, its processing and its consultation. This environment is intended to serve as reference for farmers and the community interested in using new technologies, techniques and services based on IoT that seek to improve agricultural productivity.

It has been possible to integrate all the components proposed in this research to be based on the Lambda architecture, considering the Arduino YÚN capture module as a data acquisition component, the Ubidots online database in the storage stage, the Weka

data mining tool used as the processing stage and the Java Spark web design framework for the query stage; resulting in a testing environment for IoT.

As a future work, it is intended to evaluate this IoT test environment in other application contexts where the capture of high volumes of data is required to perform a real-time processing of these, so that contexts such as education, marketing, smart cities, etc.

## References

1. Deshmane, S.: Using the Lambda Architecture on a Big Data Platform to Improve Mobile Campaign Management. Talentica Software (2015)
2. Fundación de la Innovación Bankinter: El Internet de las Cosas - En un mundo conectado de objetos inteligentes (2011)
3. IoT Simple S. A.: Agricultura inteligente (2017)
4. Jimenez, J., Hincapié, J.A., Quintero, J.: MDSD Multi-Plataforma: Más Allá de la Vista Funcional. Revista Ingenierías - Universidad de Medellín (2016)
5. Lopez, J., Chavez, J., Sánchez, A.: Modelado de una red de sensores y actuadores inalámbrica para aplicaciones en agricultura de precisión. In: Humanitarian Technology Conference (MHTC), IEEE Mexican. Puebla: IEEE (2017)
6. Ruiz, D.: Diseño de un sistema en cloud para controlar dispositivos IoT a través de Internet. Facultad de Ingeniería. Universidad Militar Nueva Granada (2016)
7. Schwartz, M.-O.: Wireless Security Camera with the Arduino Yun. Adafruit Learning System (2014)
8. Internet de las cosas (IoT) para la agricultura Colombia (2017)
9. Taffernaberry, J.C., Diedrichs, A., Pérez, C., Pecchia, M., Tabacchi, G.: Gateway 6LoWPAN para red de sensores inalámbricos. In: XLV Jornadas Argentinas de Informática e Investigación Operativa. Buenos Aires-Argentina (2016)
10. Torquica, L., Guzmán, M.: Desarrollo de Sistema de Seguridad para Automóviles con IOT y Smartphone. Facultad de Ingeniería. Bogotá: Universidad Francisco José de Caldas (2016)

# Design and Implementation of a Stand-Alone Home Energy Management System Based on Internet of Things

Javier E. Sierra, Boris Medina, Ramón Álvarez

Universidad de Sucre, Faculty of Engineering, Sincelejo, Colombia  
{javier.sierra, ramon.alvarez, boris.medina}@unisucra.edu.co

**Abstract.** Smart Grid is the most important technology in energy management, which allows to maintain a balance between supply and demand of electric energy. Traditionally, residential users demand energy incrementally, without control or management over the process, increasingly leading to high consumption and high costs in energy bills. Home Energy Management Systems (HEMS) allow the supply, optimization and automatic control of electrical and electronic appliances, distributing the load at day time, according to energy costs. In this article, we present the results of implementing a stand-alone HEMS in a house, which consists of a central node that has a database, smart meters, an autonomous control software that is fed by a mathematical model and an interconnection with Internet for management by an APPs. Internally in the house, electrical and electronic devices are adapted with Internet of Things (IoT), using Arduinos, ethernet and wifi modules, actuators and sensors. The results of the implementation show the effectiveness of HEMS in terms of cost of electricity, demand and user comfort.

**Keywords:** smart grid, internet of things, IoT, HEMS, home energy management systems.

## 1 Introduction

Electric energy currently represents quality of life, economic development and is largely responsible for many of the technological advances made by mankind [1,2]. Today it is inconceivable a community that does not enjoy the great benefits derived from the service of electric energy, such as: specialized medical assistance centers, electric transport systems, information and communication technologies, entertainment systems, appliances to perform tasks in the home, among others. It is evident that we live in an intelligent, interconnected and electro-dependent society, for which electricity is a vital element to preserve the world as we know it today [3].

The exponential growth of energy demand exceeds the implementation of conventional generation systems; this has generated alarms in the world energy sector. In addition, there is a great concern for the massification of energy sources based mainly on the use of fossil resources: coal, natural gas, oil and its derivatives. The use of fossil fuels has caused great concern about the impact on the environment, and it is therefore necessary to investigate and implement strategies aimed at saving or developing alternative sources of generation [4]. Recent research shows the possibility of reducing energy consumption by installing Home Energy Management System

(HEMS). These systems allow the active participation of the consumer in the electricity market, as well as modulate the demand curve, reducing costs, depending on a certain user profile. The user programs a consumption profile which is monitored through a local power management system, which uses intelligent sensors located at each point of consumption and software that manages a database that stores the user profiles [5].

The implementation of the energy management system for the home requires the implementation of smart sensors, relays, data network and a flexible computer platform that guarantees the management process. This management focuses on the prioritization of load consumption, in terms of costs and energy availability. The management system is fed with the consumption profiles of the residence, the patterns of saving and a table of priorities according to the season of the year.

This article proposes a new architecture of a HEMS system based on open software and hardware, as well as a mathematical model that allows to reduce the cost of energy consumption.

## 2 Basic Architectures of Energy Management Systems at Home

Home Energy Management (HEM) refers to a system that incorporates sensors to appliances through a home network [6]. HEM systems have been developed in order to measure, monitor and control energy consumption at home. Through the implementation of management software to response on demand, HEM systems enable the improvement of the performance of an electrical network. That is, HEMS software can include applications based on the profiles and preferences of residential customers as a result of the interaction between users and the electricity network. A HEM acts as a modern energy meter, being one step ahead of the low-energy consumption appliances [7]. In brief, a HEM measures, monitors and allows adjustment of energy consumption in an intelligent way, through smart meters, devices, appliances and plugs.

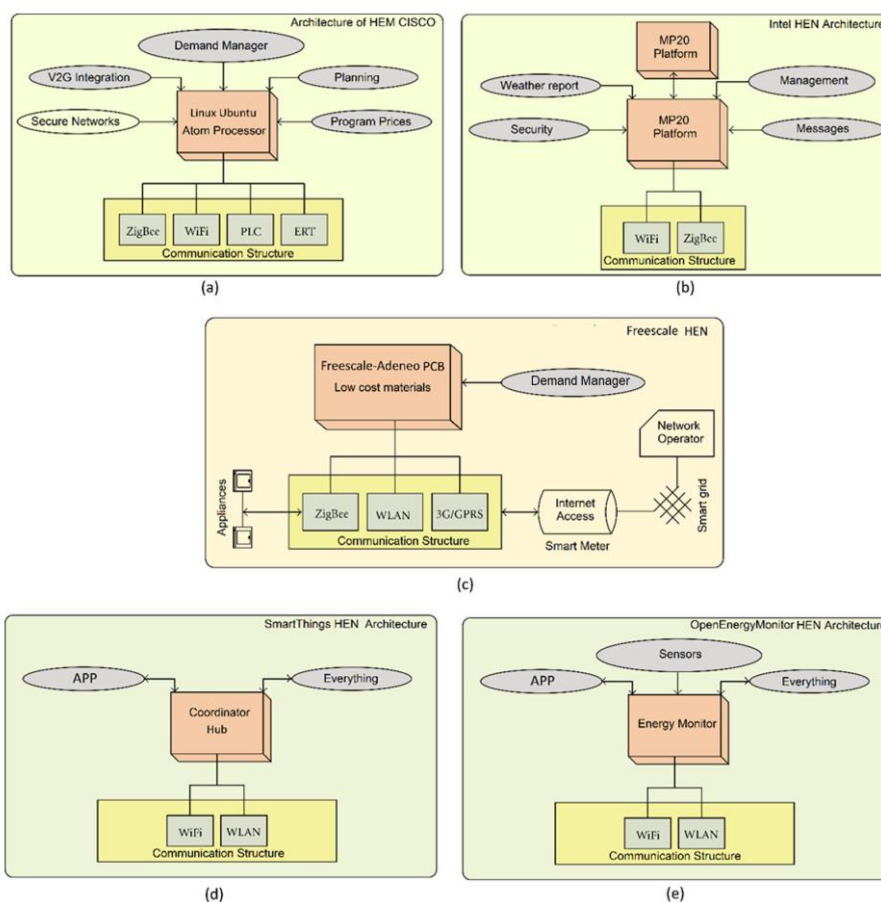
**Table 1.** Classification of articles according to the similarity in architecture

Classification	Architectural Overview
Central Controller	Central controller added to monitoring and control nodes found in appliances.
Integrated Module	Module for data collection or management, a module for devices control and a communication module
Monitoring server	Only monitoring and control devices and a domestic server
Monitoring server and gateway	Monitoring and control devices, a home server and a gateway
Hybrids	More elaborate architecture, due to the additional benefits of the HEM presented.

Architectures proposed in the literature [8] can be classified into different groups attending to several criteria related with the monitor structure, the distribution of management packages and the communication capabilities. Table 1 shows the classification of the articles according to similarities in their architecture.

Although some of the authors indicated in Table 1 present a more elaborate architecture, it is possible to establish a general architecture that includes the basic components required in a HEM. This way, a HEM basically should have:

- A home area network (HAN): It is a local residential network that interconnects devices in a house such as; sensors, smart plugs, smart thermostats and appliances allowing communication between them, either through a wireless network or a wired network.
- Monitoring and control devices: They are end devices that are responsible for monitoring and controlling the energy consumption of household appliances.
- A processor: Used for the concentration, storage and information management. The server and the database would be located in this central module.
- A gateway: Allows connection between the HEM and the outside, so that remote access via Internet can be possible.



**Fig. 1.** Basic Architectures of commercial HEMS.

In Fig. 1, five of the found architectures for defining a HEM are summarized. Fig. 1(a) is a HEM that processes in a centralized manner (Atom Processor for CISCO) the services requested by the different levels of software loaded in a platform integrated by a Linux core and external communication modules. The architecture of Fig. 1(b), differs from the previous one, by the use of a multiprocessor structure, which consists of a central processor and auxiliary processors to support the computing tasks. A practical case of such architecture is the HEM proposed by Intel using an MP20 platform. In addition, we have HEM that use a distributed structure based on a communication channel that integrates sensors, household appliances and the central processor, Fig. 1(c). A practical example of the distributed architecture is the one developed by Freescale, which is characterized by the use of low cost devices. On the other hand, they have the compact architectures, conformed by a monitor or coordinator that performs the management of household appliances and is used as interface of the user with an APP. Two practical cases of the compact architectures are the case of the SmartThing HEM and OpenEnergyMonitor, Fig. 1(d) and Fig. 1(e).

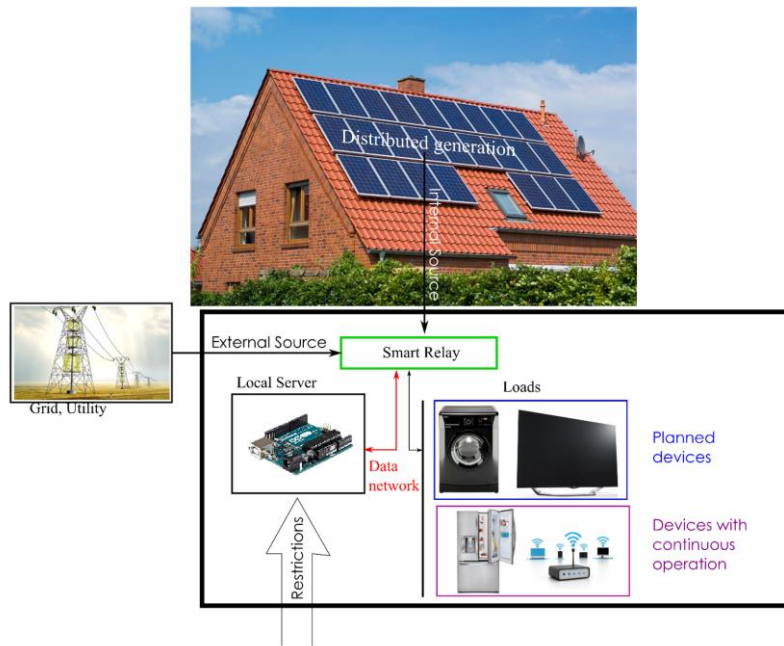


Fig. 2. Basic Architectures of proposed HEMS.

### 3 Proposed Architecture for Home Energy Management Systems

Our proposed architecture is shown in Fig. 2. The architecture employs a four-core computing system and a Graphics Processing Unit (GPU), which enables process optimization through multi-thread programming. The proposed architecture uses smart

relays, which allow load interconnection and the sensing of electrical variables. Each electrical outlet is equipped with a smart relay with sensing and transmission of electrical variables such as voltage, current and power. In addition, the proposed HEM manages continuous and programmed loads to optimize the energy demand.

The control software is implemented under the philosophy of open source, under the Linux platform. The user interface receives the information corresponding to user consumption profiles, which are due to many factors such as: the policies and regulations defined by the regulator of the electrical system, economic and socio-cultural aspects, environmental requirements and technology used (in the future these variables may be considered). It is clear that the future of energy management systems in the home is very promising, but due to the large number of aspects associated with consumption profiles, it is necessary to continue working to overcome many difficulties associated with such technology in the present.

## **4 Mathematical Model**

We consider a power system installed at the end of the user that can calculate, optimize and manage the flow and use of energy. Fig. 2 shows the system model considered where a HEMS module is connected to the smart meter and to all devices. The smart meter is a device that calculates energy consumption and communicates with the unit simultaneously. The meter receives utility pricing signals from Power Line Carrier (PLC) and communicates price values to HEMS. The smart meter, HEMS and home appliances are connected through the local area network (LAN) and share a common control channel [9].

The basic functions of the HEMS module are to collect the data of the devices, the processing of the data and to control the loads. During the data collection, the HEMS identifies the electricity consumption, calculates the price of the electric energy used in kwh and collects data relating to the priority of the customers for different appliances. Additionally, it determines the type of energy source to be used, if the one of the electric supplier or the one of an alternative source that is in home. The processing contains the analysis of the data collected and carries out the strategies for the control of the load. The planning strategies consist of modeling the system according to the mathematical model indicated in the present article. Finally, the HEMS controls the load based on the developed schemes. Communication between devices and HEMS for data collection and load control can be done with the help of existing LAN access protocols, eg Wi-Fi, Zigbee that can accommodate various communication applications [10]. In this article we consider two types of devices:

- Type 1 Planned devices: these are fully flexible devices and can be activated or programmed at a later time when the price of electricity in real time is reasonable. For example, washing machine, dishwasher and air conditioning. Let 'X' be the total number of devices that can be planned.
- Type 2 Devices with continuous operation: these are devices that have a low degree of flexibility and depend on the basic needs and the priority of the consumers. Lighting devices, computers and televisions are examples of this category. Let "Y" be the number of devices with continuous operation.

The rapid increase in energy demand forces energy companies to produce high-cost electricity, which directly affects the budget and user fees. The proposed system can significantly reduce the cost of residential consumption. The aim is to minimize the cost by programming the system devices such that demand in a given time interval does not lead to a peak in the load curve and that the operation of the devices in real time does not affect the user. Let  $T = \{t_1, t_2, t_3 \dots t_N\}$  be the set of  $N$  scheduling time slots with  $t_n$  denoting the  $n$ -th slot. Generally, the behavior of energy use is random and has scheduling ranges in which a higher consumption occurs. We define the set of schedulable devices  $S = \{a_1, a_2, a_3 \dots A\}$  and the set of real-time devices  $R = \{b_1, b_2, b_3 \dots B\}$ . A series of real-time and schedulable devices may be active at each time interval of the set  $T$ .

We define a binary variable  $v_{i,n}$  such that:

$$v_{i,n} = \begin{cases} 1 & \text{if } i\text{th devices is ON in time } t_n \quad \forall i = 1 \dots X, n = 1 \dots N, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Therefore, the number of planned devices (type 1) that are activated in the time slot  $t_n$  can be represented as:

$$\gamma_{ON}^n = \sum_{i=1}^X (v_{i,n}), \quad \forall n. \quad (2)$$

We define  $z_{j,n}$  a binary variable for devices with continuous operation (type 2):

$$z_{j,n} = \begin{cases} 1 & \text{if } j\text{th devices is ON in time } t_n \quad \forall i = 1 \dots Y, n = 1 \dots N. \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Thus, in a given time interval  $t_n$ , devices with continuous operation are:

$$\Delta_{ON}^n = \sum_{j=1}^Y (z_{j,n}), \quad \forall n. \quad (4)$$

The proposed model has the following restrictions:

To ensure that demand at peak hours does not increase greatly, the energy consumed by the combination of devices with continuous operation and the planned devices in any time interval must be kept under a target value  $E$ . Therefore, we have:

$$\sum_{i=1}^X \varphi_i^n + \sum_{j=1}^Y \omega_j^n \leq E, \quad \forall n, \quad (5)$$

with  $\varphi_i^n = (P_{i,n})(v_{i,n})$  and  $\omega_j^n = (Q_{j,n})(z_{j,n})$  where  $P_{i,n}$  and  $Q_{j,n}$  are the powers consumed in time slot  $t_n$  by the  $i$ th planned device and the  $j$ th devices with continuous operation, respectively.

Ideally the devices in set  $Y$  should be turned ON all time due to lower degree of flexibility:

$$\sum_{j=1}^Y (z_{j,n}) = |Y|, \quad \forall n, \quad (6)$$

$$\sum_{n=1}^N (z_{j,n}) = N, \quad \forall j, \quad (7)$$

where the operator  $|\cdot|$  denotes the cardinality of the set. However, for a greater number of devices with continuous operation, it may be impossible to accommodate all in each time interval. Therefore, the above restrictions are reformulated as:

$$\Delta_{ON}^n = Y', \quad \forall n, \quad (8)$$

$$0 \leq \sum_{n=1}^N (z_{j,n}) = N, \quad \forall j, \quad (9)$$

where  $Y' = Y$  if  $\sum_{j=1}^Y w_j^n \leq E$  and  $Y' \subset Y$  if  $\sum_{j=1}^Y w_j^n > E$ . The above expressions ensure that at least one devices with continuous operation is turned ON and if a device is turned on it will remain active for the entire time. Planned devices have high operational flexibility. If in a given slot, the real time devices demand more than  $E$  no device from the set  $S$  will be scheduled. On the other hand, for a very limited requirement from set  $R$ , all the 'X' devices can enjoy the turned ON status. Thus:

$$\varphi_{ON}^n = X', \quad \forall n, \quad (10)$$

$$\sum_{n=1}^N (v_{i,n}) \leq N, \quad \forall i \quad (11)$$

where  $X' = X$  if  $\sum_{i=1}^X \varphi_i^n \leq E - \sum_{j=1}^Y w_j^n$  and  $X' \subset X$  in case of  $\sum_{i=1}^X \varphi_i^n > E - \sum_{j=1}^Y w_j^n$ . Unlike the previous case, the latter equation shows that depending on the required operating time, a particular device could be programmed only for a fraction of the full-time window.

### Objective function

Let  $C_n$  represent the per unit cost at time  $t_n$ . Thus, during  $n$ -th slot the cost of a planned devices and devices with continuous operation are  $\gamma_{planned}^n = P_{i,n} C_n$  and  $\gamma_{cont}^n = Q_{j,n} C_n$ , respectively. Our target is to minimize the sum cost over all the scheduling hours such that no violation occurs for any stated constraint. The optimization problem can be defined mathematically as:

$$\min_{v_{i,n}, z_{j,n}} \sum_{t=1}^N \left( \sum_{i=1}^X \Gamma_{i,n}^{planned} (P_{i,n}, \varphi_i^n, \gamma_{planned}^n) + \sum_{j=1}^Y \Gamma_{j,n}^{cont} (Q_{j,n}, w_j^n, \gamma_{cont}^n) \right), \quad (12)$$

with the constraints shown in equations (3), (6), (7), (8),(9) and  $v_{i,n} \in \{0,1\}, z_{j,n} \in \{0,1\} \forall i, n$ . The two cost functions  $\Gamma_{i,n}^{planned}$  and  $\Gamma_{j,n}^{cont}$  represent the cost of  $i$ -th planned devices if it is scheduled in  $n$ -th time slot and  $j$ -th devices with continuous operations when it is scheduled to turn ON in  $n$ -th tiem slot, respectively, and are given:

$$\Gamma_{i,n}^{planned} = \frac{\varphi_i^n \gamma_{planned}^n}{P_{i,n}}, \quad (13)$$

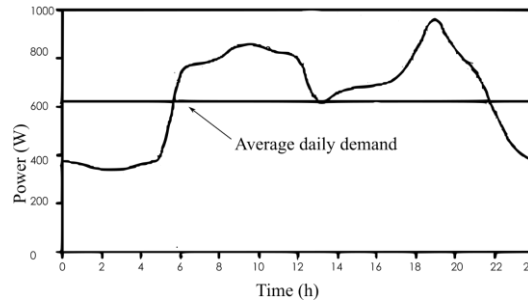
$$\Gamma_{j,n}^{cont} = \frac{w_i^n \gamma_{cont}^n}{Q_{i,n}}. \quad (14)$$

The above optimization aims to find the total of  $N(X + Y)$  variables, i.e., the optimum values of  $v_{i,n}, \forall i, \forall n$  and  $z_{j,n}, \forall j, \forall n$ , which provide the minimum possible cost keeping total demand in each hour under a predefined limit.

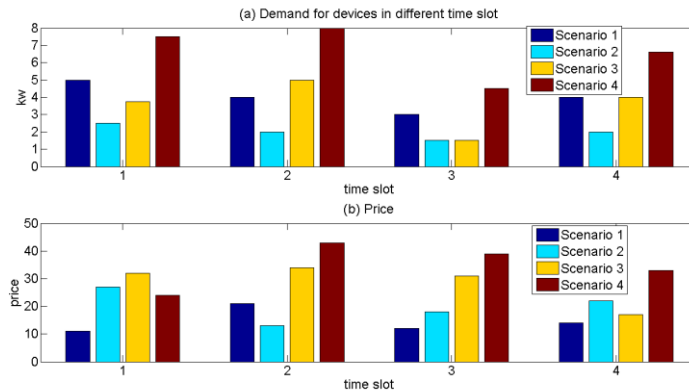
The formulated problem is a mixed binary integer programming problem and has a high computational complexity to find the optimal solution.

## 5 Results

We simulate in the CPLEX software the mathematical model to verify the effectiveness of our proposed architecture. The demand of the users varies during the course of the day, in Fig. 3 shows a typical behavior of a residential user. Given this, we simulate for several time slot ( $T = \{t_1, t_2, t_3, t_4\}$ ) in 4 different scenarios. For the purpose of the simulation we consider that the user has fixed ten applications, but the demands of several devices may vary over the course of the day.



**Fig 3.** Typical Residential Demand Response.



**Fig 4.** Demand and Cost for devices in different time slot.

There are 8 devices connected to the network for the simulation, of which two of them are operating in continuous time, ie always remain ON. Fig. 4a shows the demands required for each time slot in the four simulated scenarios. We have considered the TOU pricing model [11,12], that will assign a different cost ( $p_i$ ) for each  $t_i$ , where  $p_i$  represents the price per unit. Each set has a unique cost in a given time interval, although it varies in different time intervals due to the TOU price model. For example, the cost of each scenario in scheduling schedules is shown in Fig. 4b.

Fig. 5 presents the total cost of scheduling the demands for different load scenarios, in each of the time slots. It is observed that the mathematical model determines the cost for each scenario, identifying the elements that must be programmed in the required times. It is necessary to propose heuristic techniques that seek to solve the problem raised in an optimum time.

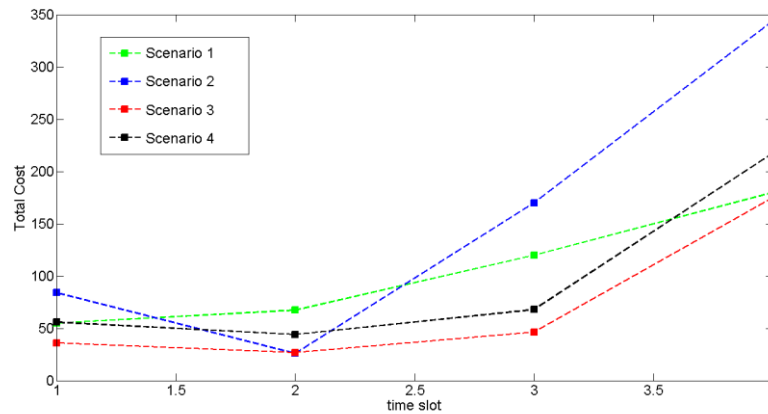


Fig. 5. Total cost in different time slot.

## 6 Conclusions

This article proposes a new architecture for domestic use that uses a Home Energy Management System (HEMS). The proposed architecture consists of a server in the house that collects the data of consumption and of the generation of energy, analyzes them through the proposed algorithm and controls the use of energy to minimize the cost of energy. The architecture, as well as the mathematical model evaluated show that it is possible to minimize the cost of energy consumption.

## References

1. Gyorvari, B., Vokony, I.: Technical issues of solar- and gas engine based MicroGrids: Assessment on feasibility by using present technologies. In: 2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe) 2017, pp. 1–6 (2017)
2. U.S. Energy Information Administration: EIA's Annual Energy Outlook (2017)

3. Basit, A., Sidhu, G.A.S., Mahmood, A., Gao, F.: Efficient and Autonomous Energy Management Techniques for the Future Smart Homes. *IEEE Trans. Smart Grid*, pp. 1–10 (2015)
4. Langhammer, N, Kays, R.: Performance Evaluation of Wireless Home Automation Networks in Indoor Scenarios. *IEEE Trans Smart Grid*. 3(4), 52–61 (2012)
5. Han, J., Choi, C., Park, W., Lee, I., Kim, S.: Smart home energy management system including renewable energy based on ZigBee and PLC. *IEEE Trans. Consum. Electron.* 60(2), 198–202 (2014)
6. Choi C.S., Park W.K., Han J.S., Lee I.W.: The architecture and implementation of proactive Green Home Energy Management System. In: 2010 International Conference on Information and Communication Technology Convergence, pp. 457–458 (2010)
7. D'Alessandro, S., Tonello, A.M., Monacchi, A., Elmenreich, W.: Home energy management systems: Design guidelines for the communication infrastructure. In: *ENERGYCON 2014 - IEEE International Energy Conference*, pp. 805–812 (2014)
8. De la Cruz, Z., Alvarez-Lopez, R., Fernandez-Vilas, A., Díaz-Redondo, R.: Tendencia tecnológica de los sistemas de administración de energía eléctrica en el hogar: revisión bibliográfica. *DYNA Energía y Sostenibilidad* 5 (2016)
9. Zhou, B., *et al.*: Smart home energy management systems: Concept, configurations, and scheduling strategies. *Renew. Sustain. Energy Rev.* 61, 30–40 (2016)
10. Cheah, P.H., Zhang, R., Gooi, H.B., Yu, H., Foo, M.K.: Consumer energy portal and home energy management system for smart grid applications. In: 2012 10th International Power & Energy Conference (IPEC), pp. 407–411 (2012)
11. Huang, Y, Tian, H, Wang, L.: Demand response for home energy management system. *Int J Electr Power Energy Syst.* 73, 448–455 (2015)
12. Mehta, R., Srinivasan, D., Verma, P.: Intelligent appliance control algorithm for optimizing user energy demand in smart homes. In: 2017 IEEE Congress on Evolutionary Computation (CEC), pp. 1255–1262 (2017)

# Analysis of Feature Sets for Malware Family Classification

Jesús Javier Reyes Torres, Eleazar Aguirre Anaya, Ricardo Menchaca Méndez,  
Nareli Cruz Cortés, David Alejandro Robles Ramírez

Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico  
{eaguirre, ric, nareli}@cic.ipn.mx  
{b151223, b151224}@sagitario.cic.ipn.mx

**Abstract.** Malware families are evolving constantly, in order to evade the detection mechanisms, the authors modify the order of functions, add random useless code and add new features, it has been observed that these new variations share common characteristics with previous versions, these existing patterns in the malware allows us to generate characteristics to describe it, for later use in a machine learning algorithm. In this paper, we analyze different feature sets extracted from malicious portable executables which are used as input to a machine learning algorithm, these features are extracted using n-grams, and are used in three classification models: Logistic Regression, Random Forest and Support Vector Machines.

**Keywords:** malware classification, N-grams, portable executable, machine learning, logistic regression, random forest, support vector machine.

## 1 Introduction

Malicious software currently represents a computational security problem, so it is important to recognize the prevalence and continuing growth of malicious software. In the year 2015, 16.7 million of new variants of malware were found, according to Symantec [1]. Most of the new malicious software is designed to evade the anti-malware systems which use signature based methods to detect it. The authors modify the order of the functions, add random useless code, obfuscation techniques as packing in order to evade the detection methods.

The malicious software analysis techniques are classified into dynamic and static approach. In dynamic analysis, malware's information is collected from the operating system at runtime, such information can be system calls, network access and files, in this approach it is hard to simulate appropriate conditions in which the malicious software can execute its malicious functions, and at the beginning we do not know the time period needed to observe the malicious activity of the program. In static analysis, the information is gathered from the executable file without executing it, this information represents the expected behavior, in this approach features are extracted by an analyst, these features can be from the disassemble code or from the hexadecimal view of the binary.

In this paper we address the classification, using static features extracted from nine malicious software families obtained from a Microsoft database. Our approach is a non-signature based method in which we use machine learning to generate a classification model to classify a given sample in one of the nine families used to generate the model.

## 1.1 Scope

The most commonly used operating system is the Microsoft Windows operating system, this make the operating system an attractive target for malware authors. The principal executable format in Windows is the Portable Executable format (PE), which is a Microsoft standard. In our work, all the samples are malicious software written in the Portable Executable format, some of them for 32-bit systems and others for 64-bit systems.

## 2 Related Work

Different approaches for malicious software detection have been presented for years, at first F. Cohen [2] showed that in general the problem of virus detection is undecidable. In [3] Harris and Miller establish that the characteristics extracted through the analysis of the binary code provide information about the content and structure, for example instructions, basic blocks, functions and modules.

Machine learning for malware detection has been widely used, Siddiqui *et al.* extract variable length instructions sequences that can identify trojans from clean programs using data mining techniques [4]. Schultz *et al.* in [5] present a data mining framework to detect new malicious executables, they use different algorithms to generate classification models, the Multi-Naïve Bayes method was reported as the one with the best accuracy and detection rate over unknown programs with the value of 97.76%.

In [6] different patterns were used to detect the presence of malicious content in executable files. The analysis is made principally taking in consideration the bytecode as in [7] where they compute statistical and information-theoretic features in a block-wise manner to quantify the byte-level file content, and in [8] the authors use a static analysis methodology for representing malicious codes, their framework seeks to acquire the most important files, benign and malicious, in order to improve classifier performance.

N-grams is a way to represent the malicious software content, it consists in generate substrings with length  $n$  from a larger string, since n-gram overlap, they do not capture just statistics about substrings of length  $n$ , but they implicitly capture frequencies of longer substrings as well [9], in [10] Santos *et al.* demonstrate that a n-gram-based methodology signatures can achieve detection of new or unknown malicious software, Abou-Assaleh *et al.* [9] demonstrates that applying the CNG method based on byte n-gram analysis good results can be achieved, in [11] the authors use the n-grams to develop an automatic malware categorization system (AMCS) by observing the common characteristics shared by different malicious software families.

The operational code (OpCode) also has been used as static information to represent malicious software, in [12] O’Kane *et al.* use OpCode to detect encrypted malware using SVM, Santos *et al.* in [13] propose a method to detect unknown malware families, using

the frequency of appearance of OpCode sequences as a base, and in [14] Bashari Rad et al. use static analysis to generate histograms of machine instructions frequency to be used as a features to classify the obfuscated version of metamorphic viruses.

Narayanan *et al.* in [15] obtained suggestions that every malware software belonging to a family has a distinct pattern, these patterns are quite similar between a family and distinguishable across other families.

In [16] Srakaew *et al.* compare two malware classification methods using data mining, they use two different types of features: statistical features and abstract assembly features, they observed that the abstract assembly approach is more promising, giving high accuracy with less complicated model.

### 3 Data Set

The data set was obtained from a Microsoft repository, this data set contain nine different malware families which are Ramnit, Lollipop, Kelihos\_ver3, Vundo, Simda, Tracur, Kelihos\_ver1, Obfuscator.ACY and Gatak, the data set is unbalanced as can be observed in Fig. 1.

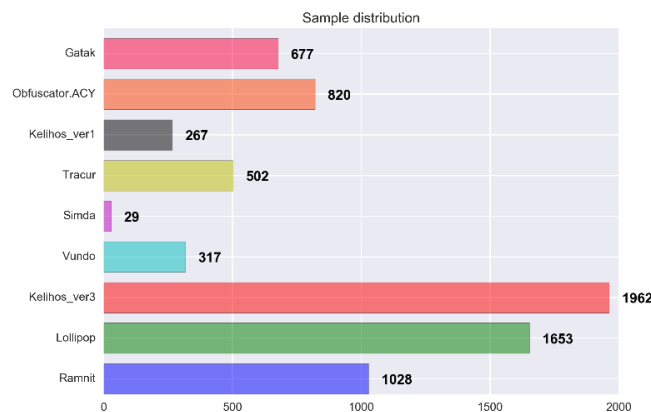


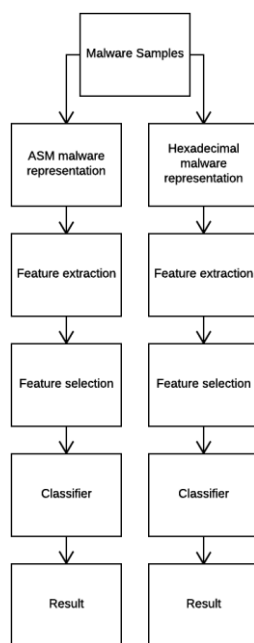
Fig. 1. Sample distribution.

### 4 Methodology

In this section, we describe the malicious software classification process, in our work we use static features generated from the malicious samples. In general, there are three distinct stages when a machine learning malicious software detection approach is used, these sections are: Feature extraction, sometimes feature selection is applied in order to reduce the dimensionality of the file's representation, and the generation of a classifier model using a machine learning algorithm.

The flow followed in this work is shown in Fig. 2. in each stage, different methods are used and a different data representation is obtained. In the malware analysis process, the analyst mostly of the time only have the malicious executable to start with, is very

improbable to have the source code of a malware sample. So, it is convenient to transform the executable into a more manageable representation, in this work the ASM and hexadecimal view are used as the representation of the malicious files.



**Fig. 2.** General framework of malware classification system.

These new representations are processed to extract the information with which we want to work with, from the hexadecimal representation we extract only the hexadecimal code, the memory addresses are not taken into account.

```

00000000 0000 0001 0001 1010 0010 0001 0004 0128
00000010 0000 0016 0000 0028 0000 0010 0000 0020
00000020 0000 0001 0004 0000 0000 0000 0000 0000
00000030 0000 0000 0000 0010 0000 0000 0000 0204
00000040 0004 8384 0084 c7c8 00c8 4748 0048 e8e9
00000050 00e9 6a69 0069 a8a9 00a9 2828 0028 fdfe
00000060 00fc 1819 0019 9898 0098 d9d8 00d8 5857
00000070 0057 7b7a 007a bab9 00b9 3a3c 003c 8888
00000080 8888 8888 8888 8888 288e be88 8888 8888
00000090 3b83 5788 8888 8888 7667 778e 8828 8888
000000a0 d61f 7abd 8818 8888 467c 585f 8814 8188
000000b0 8b06 e8f7 88aa 8388 8b3b 88f3 88bd e988
000000c0 8a18 880c e841 c988 b328 6871 688e 958b
000000d0 a948 5862 5884 7e81 3788 1ab4 5a84 3eec
000000e0 3d86 dcb8 5cbb 8888 8888 8888 8888 8888
000000f0 8888 8888 8888 8888 8888 8888 8888 0000
00001000 0000 0000 0000 0000 0000 0000 0000 0000
*
00001300 0000 0000 0000 0000 0000 0000 0000
000013e
  
```

**Fig. 3.** Hexadecimal view from an executable file.

From the ASM representation we only use the OpCode field, ignoring the label field, the operand field and comment field:

[Campo de etiqueta] [Campo de Opcode [Campo del operando]] [Campo de comentarios]

The complete set of mnemonics present in the samples is not used, instead we use a reduced set of mnemonics, considered to be good in describe the malware samples. These instructions were selected using previous works in this topic.

**Table 1.** Reduced set of mnemonics.

and	fstcs	test	nop
int	setle	lea	shld
jnz	xor	jz	jb
fild	sub	jmp	std
imul	fdvip	mov	sbb
pop	retn	movzx	setb
loopd	ja	bt	push
jnb	add	lods	dec
pushf	call	rdtsc	rep
inc	adc	je	cmp

## 5 Feature Extraction

The features are created with n-grams applied to the final representation of the sample, the final executable representation with the hexadecimal representation is shown in Figure 5. and the final executable representation with the ASM representation is shown in Fig. 6.

```
4d5a90000300000004000000ffff0000b800000000000000400000000
0000000000000000000000000000000000000000000000000000000
0000000800000000e1fba0e00b409cd21b8014ccd21546869732070726
f6772616d2063616e6e6f742062652072756e20696e20444f53206d6f
64652e0d0d0a240000000000000504500004c010300de91895700000
0000000000e00022000b013000004e0000000a00000000000b26d00
00002000000080000000004000002000000002000004000000000000
004000000000000000c0000000020000ed5c00000200608500001000
0010000000001000001000000000000010000000000000000000006
06d00004f0000000800000980600000000000000000000000000000
```

**Fig 4.** Hexadecimal representation of an executable file.

To these representation, the n-gram extraction method is applied, for the case of the hexadecimal we consider each character of the hexadecimal alphabet as a unit.

```
xor retn int int push lea sub mov push push
push mov mov call mov lea push push call mov
push call push call push push call push call
mov mov imul mov mov mov add mov mov mov lea
mov inc test jnz mov sub mov add mov mov mov
mov test jz mov cmp jz cmp jz cmp jz xor jmp
mov mov mov add mov add mov add mov mov mov
mov sub add mov mov mov mov add mov mov mov
mov mov test jz mov mov xor pop pop pop add
mov pop retn push push push push mov call mov
```

**Fig 5.** ASM representation of an executable file.

In Fig. 7. can be observed the process to generate the n-grams from a string with a value of  $n$  equal to 4, the number of n-grams generated from a string with length  $L$  can be calculated as follows:

$$Gr = L - (n - 1), \quad (1)$$

where  $n$  is the length of the sub-strings.

```
F0327D609548D06208804B67B44A21DC
F032
0327
327D
.....
A21D
21DC
```

**Fig. 6.** N-grams generation process in the hexadecimal representation.

For the case of the ASM representation we consider each mnemonic as a unit, in Fig. 8. can be observed the process to generate the n-grams in this representation.

```
xor retn int int push lea sub mov push push
xor retn
  retn int
    int int
      .....
        mov push
          push push
```

**Fig. 7.** N-gram generation process in the ASM representation.

After processing the samples into n-grams, the set of samples  $S$  can be represented as a matrix where each column represents a n-gram term and each row represent a sample  $s \in S$ . As we are using the supervised learning approach we also have to include the label of each sample, in our case we have nine levels, each one representing a malware family.

As can be observed in Fig. 9. between all the  $k$  samples which compose  $S = \{s_1, s_2, \dots, s_k\}$  we generate a set of terms (n-grams)  $T = \{t_1, t_2, \dots, t_m\}$  which is composed for all the unique strings whit length  $n$  generated by the extraction over all

the sample space, each cell in the terms space is filled with the normalized term frequency, showed in Equation 2:

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}. \quad (2)$$

Here we divide the frequency of a term  $t$  in a determined document  $d$  by the maximum frequency of some term  $w$  in the same document.

Sample label	Terms (N-gram)					
$s_1$	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_m$
$s_2$	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_m$
$s_3$	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_m$
$s_4$	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_m$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$s_k$	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_m$

Fig. 8. Samples representation matrix.

## 6 Feature Selection

With the n-grams a large number of features is generated, this number may vary according to the unique number of sub-strings that can be found in the extraction process over the sample space, this number depends in the alphabet used to represent the data, in our case for the hexadecimal representation the alphabet is  $A_H = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$  which are the characters used to represent hexadecimal digits, and for the case of the ASM representation, the alphabet  $A_{ASM}$  is composed by the forty mnemonics in Table 1.

The number of substrings which represent the features in our machine learning approach has as an upper bound all the possible unique strings whit length  $n$  that can be generated with our alphabet.

$$UB = (Alphabet)^n. \quad (3)$$

As we are using machine learning to approximate the functional relationship  $f()$  between an input  $X = \{x_1, x_2, \dots, x_M\}$  and an output  $Y$ , based in a tuple relation  $\{X_i, Y_i\}$ ,  $i = 1, \dots, N$ , sometimes the output  $Y$  is not determined by the complete set of the input features  $\{x_1, x_2, \dots, x_M\}$ , instead, it is decided only by a subset of them  $\{x_{(1)}, x_{(2)}, \dots, x_{(m)}\}$ , where  $m < M$ .

The irrelevant features may lead in increase the computational cost and overfitting, a rule of thumb about the relation between the number of samples and the features to describe them, this rule claim that the numbers of samples needed to make a good generalization has to be more or equal to ten times the number of effective features.

$$N \geq 10(effective\ features). \quad (4)$$

Our sample space is composed of 7255 malicious software samples, and if for instance we generate the n-grams with a value of  $n = 4$  over the hexadecimal

representation, the upper bound number of features is 65536, which is a large number of features compared with the number of samples.

## 7 Feature Sets

We generate different feature sets and use them as an input of a machine learning algorithms to generate different classification models. The different feature sets generated are shown below:

- Features generated using hexadecimal representation and the n-gram method with a value of  $n = 2$ .
- Features generated using hexadecimal representation and the n-gram method with a value of  $n = 4$ .
- Features generated using opcode representation and the n-gram method with a value of  $n = 2$ .

Also, a feature vector is generated using those generated by the hexadecimal representation with a value of  $n = 2$  and opcode representation with a value of  $n = 2$ . The number of features generated with the different representations are shown in Table 2.

**Table 2.** Number of features generated.

Features	Number of features
Hexadecimal $n = 2$ (D1)	257
Hexadecimal $n = 4$ (D4)	65537
Opcode $n = 2$ (D2)	1138
Hexadecimal $n = 2$ and Opcode $n = 2$ (D6)	1394

Taking into account the rule of thumb presented in (4) in three cases this is not fulfilled, so it was proposed to use feature selection to reduce the dimensions of those vectors.

The feature selection used in this work is the univariate feature selection, using the  $\chi^2$  metric which is used to test the independence of two events, the method examines each feature independently to determine the strength of the relationship of the feature with the response variable.

**Table 3.** Number of features generated by applying feature selection.

Feature	Number of features
Hexadecimal $n = 4$ (D5)	725
Opcode $n = 2$ (D3)	725
Hexadecimal $n = 2$ and Opcode $n = 2$ (D7)	725

In Table 3 is shown the number of features generated for each representation taking into consideration that we only have 7255 samples to use in the model generation. The

algorithms in which we prove those different feature sets are Logistic Regression, Random Forest, K Nearest Neighbors and Support Vector Machines.

Important parameters of the machine learning algorithms:

- Logistic regression: L2 for penalty, balanced class weight, primal formulation.
- Random Forest: ten estimators, gini impurity, two as minimum sample splits, one sample required to be a leaf node, bootstrap samples to build a tree, balanced class weight.
- K nearest neighbors: three neighbors, minkowski metric with  $p=2$ .
- Support Vector Machine: L2 for penalty, square of the hinge loss, error term equal 10, dual problem of optimization, radial kernel, gamma equal to 0.5.

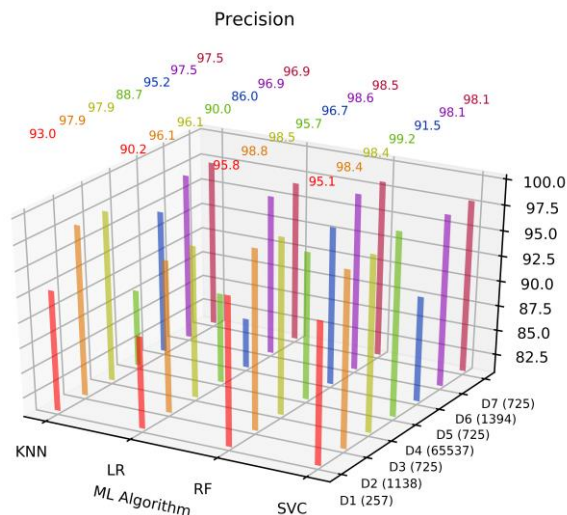
We used the cross-validation of K-Fold as validation method, as shown in Fig. 1, the samples representing each family of malware are unbalanced so we use the stratified method of the method, to maintain the representativeness of each family when the model is generated.

## 8 Experiments and Results

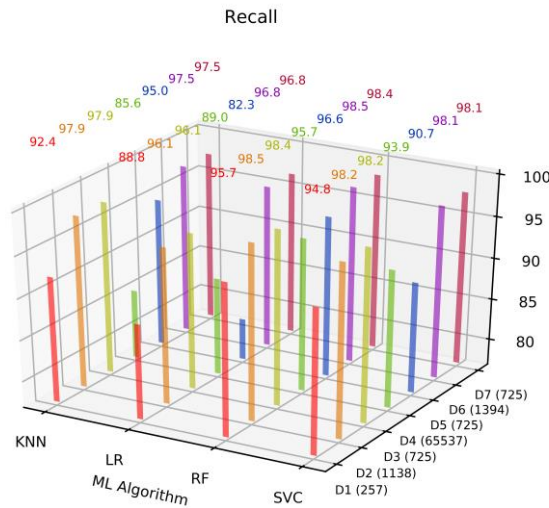
The experiments were made in a computer system with the following characteristic: processor Intel Core i5 at 2.8 GHz, 16 Gb of RAM memory and macOS Sierra as operative system. The software used were programmed in python 2.7.

The metrics used to analyze the results were Precision, Recall and F1-score, the reason why we do not use accuracy is that our database is unbalanced and we have to make sure that the family with the less number of samples (Simda) is detected by the model generated.

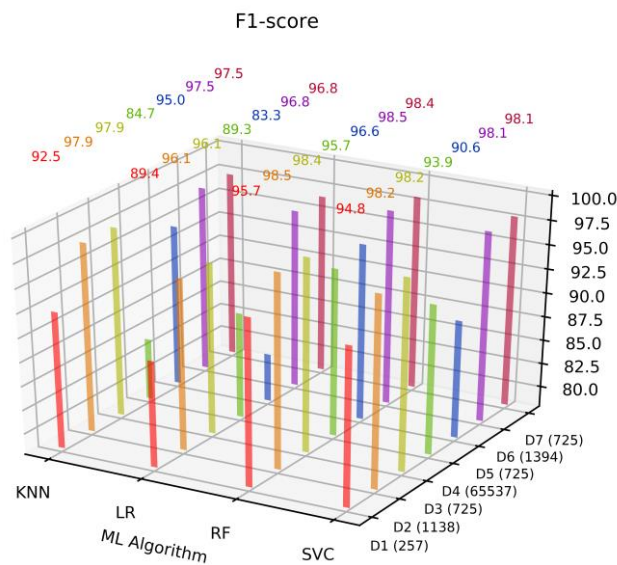
Below are the results obtained for each feature set with the different machine learning algorithms using K-Fold Cross validation with a value of  $K = 10$ .



**Fig. 9.** Precision metric for all the feature sets.



**Fig 10.** Recall metric for all the feature sets.



**Fig. 11.** F1-score metric for all the feature sets.

In the state of the art the accuracy is the metric used to evaluate the model performance, below is shown a table with some works, and their performance.

It is possible to make a comparison but it is not strict, because the metrics are not the same, we prefer not to use the accuracy due to the characteristics of the database, in addition to that in our dataset there are samples that present obfuscation. Even so the result that we obtained is competent with those mentioned in table 4.

**Table 4.** Results obtained in the state of the art.

	Approach	Results
N-gram-based detection of new malicious code	Using n-grams with a value of n from 1 to 10. CNG classification method.	Maximum accuracy of 98%
Pattern based feature selection	Using 4-gram and selecting features with PCA. Support Vector Machine.	Accuracy of 97.7%
Malware detection using statistical analysis of byte-level file content	Features extraction using statistical and information theory features. Boosted J48 algorithm.	Area Under the curve of 96.2%
Proposal of n-gram based algorithm for malware classification	Using 4-gram and the top 60 features. Similarity function to calculate.	Accuracy of 98%
Selecting features to classify malware	Using features from metadata in the PE header, data about 10 sections and all import and exports. J48 algorithm.	Accuracy of 98%

## 9 Discussion

In all the feature sets the algorithm which the best performance is obtained is Random Forest, between them can be observed that using the features generated by the opcode representation with a value of  $n = 2$  (D2) and the features generated using the hexadecimal representation with a value of  $n = 2$  plus the generated using the opcode representation with a value of  $n = 2$  (D6) highest performance is obtained, in both cases the number of features do not fulfil the rule of thumb presented in (4) so it is possible that we are overfitting the model to the data, if we look at the results obtained using the same feature sets but with the feature selection applied (D3 and D7) it can be observed that the values are not very different and It is possible that we are reducing overfitting in order to make better predictions outside the training/validation dataset.

Generally, the most difficult family to model is the is Obfuscator.ACY, the detection rate of this family in the two models with highest performance is very close to hundred percent, in the case when the feature selection was applied to the features generated, the number of samples well classified decrease at most by six percent.

Another factor taken into consideration is that the family Simda has very few samples, in the four models generated using Random Forest one third of the samples belonging to this family were classified correctly. With the four sets mentioned above also were obtained the best results for the other three algorithms compared with the remaining feature sets.

The worst results were obtained using the features generated from the hexadecimal representation with a value of  $n=4$  for the n-grams, the performance is improved by using these features applying feature selection in K Nearest Neighbor and Random Forest algorithms, but it gets worse with Logistic Regression and Support Vector Machines.

## 10 Conclusions

As can be observed in the graphs some of the feature sets are good to describe our database, the features with which the best performance is achieved are they generated using the opcode representation with a value of  $n=2$  and the case in which to these features are aggregated the features generated using the hexadecimal representation with a value of  $n=2$ , applying feature selection to both cases we achieve almost the same results, the algorithms in which the best results are obtained are Random Forest and Support Vector Machines, that can be because in Random Forest the number of features can be large, and in SVM due to the regularization parameter the algorithm is more resistant to overfitting.

In the algorithms KNN and Logistic regression we observe a improvement when the features generated with the opcode representation is used compared with these obtained using the features generated with the hexadecimal representation, so it can be say that the opcode is better to describe our data.

**Acknowledgements.** The authors thank the Instituto Politécnico Nacional and CONACYT for their support in the realization of this work.

## References

1. Symantec reports. <https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf> [Last accessed: 11 08 2017]
2. Cohen, F.: Computer viruses: theory and experiments. *Computers & security* 6(1), 22–35 (1987)
3. Harris, L.C., Miller, B. P.: Practical analysis of stripped binary code. *ACM SIGARCH Computer Architecture News* 33(5), 63–68 (2005)
4. Siddiqui, M., Wang, M.C., Lee, J.: A survey of data mining techniques for malware detection using file features. In: *Proceedings of the 46th annual southeast regional conference on xx*, pp. 509–510, ACM (2008)
5. Schultz, M.G., Eskin, E., Zadok, F., Stolfo, S.J.: Data mining methods for detection of new malicious executables. In: *Security and Privacy 2001 (S&P 2001)*, IEEE Symposium on, pp. 38–49 (2001)
6. Liangboonprakong, C., Sornil, O.: Classification of malware families based on n-grams sequential pattern features. In: *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on.*, pp. 777–782. IEEE (2013)
7. Tabish, S.M., Shafiq, M.Z., Farooq, M.: Malware detection using statistical analysis of byte-level file content. In: *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatic*, pp. 23–31, ACM (2009)
8. Shabtai, A., Moskovitch, R., Feher, C., Dolev, S., Elovici, Y.: Detecting unknown malicious code by applying classification techniques on opcode patterns. *Security Informatics* 1(1), 1 (2012)
9. Abou-Assaleh, T., Cercone, N., Keselj, V., Sweidan, R.: N-gram-based detection of new malicious code. In: *Computer Software and Applications Conference (COMPSAC 2004). Proceedings of the 28th Annual International*, vol. 2, pp. 41–42. IEEE (2004)
10. Santos, I., Peña, Y.K., Devesa, J., Bringas, P.G.: N-grams-based File Signatures for Malware Detection. *ICEIS* 2(9), 317–320 (2009)

11. Ye, Y., Li, T., Chen, Y., Jiang, Q.: Automatic malware categorization using cluster ensemble. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 95–104 (2010)
12. O’Kane, P., Sezer, S., McLaughlin, K.: N-gram density based malware detection. In: Computer Applications & Research (WSCAR), 2014 World Symposium on, pp. 1–6. IEEE (2014)
13. Santos, I., Brezo, F., Nieves, J., Peña, Y.K., Sanz, B., Laorden, C., Bringas, P.G.: Idea: Opcode-sequence-based malware detection. In: International Symposium on Engineering Secure Software and Systems. Springer, Berlin, pp. 35–43, Heidelberg (2010)
14. Rad, B.B., Masrom, M., Ibrahim, S.: Opcodes histogram for classifying metamorphic portable executables malware. In: e-Learning and e-Technologies in Education (ICEEE), 2012 International Conference on, pp. 209–213. IEEE (2012)
15. Narayanan, A., Yang, L., Chen, L., Jinliang, L.: Adaptive and scalable android malware detection through online learning. In: Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 2484–2491. IEEE (2016)
16. Piyantcharatsr, S.S.W., Adulkasem, S., Chantrapornchai, C.: On the Comparison of Malware Detection Methods Using Data Mining with Two Feature Sets. *International Journal of Security and Its Applications* 9(3), 293–318 (2015)



# Ecosystems of Collaborative Learning in Educational Technological Mediations: Case Analysis

Carolina Burbano Gonzalez<sup>1</sup>, Clara Burbano Gonzalez<sup>2</sup>, Katerine Márceles Villalba<sup>2</sup>

<sup>1</sup> Universidad Cooperativa de Colombia, Colombia.

<sup>2</sup> Corporación Universitaria Comfacaucá, Popayán, Colombia

<sup>1</sup> caritoburg@yahoo.com,

<sup>2</sup> {cburbano, kmarceles}@Unicomfacaucá.edu.co

**Abstract.** The current educational scenarios recognize the importance of interactions with each other in the acquisition of rhizomes of knowledge, knowing how to work in groups to achieve common goals appears as a transversal competence in all levels of education. At the same time, the digital technologies are in processes of expansion and generalization in the educational systems, allowing the communication between students and teachers. Hence it is necessary to think of new pedagogical practices that foster processes of knowledge, dynamising practices thought from the encounter in otherness and interactions with a multi located thought or group cognition; To bidirectional relationships between systems, virtual and human educational environments. It is necessary to take an epistemological, cultural, political, economic turn in which one starts from the co - construction of knowledge from experiences. Therefore, when talking about knowledge it is important to emphasize that science, technology and society has taken a leading role in the educational field in this XXI century, requires the management and appropriation of the same in the process of teaching and learning as a media resource And as a strategy of knowledge dialogue; It is through these resources that different spaces and times are made available for research, knowledge, and integration of models and educational strategies that lead to social and educational interpersonal and intrapersonal encounters.

**Keywords:** educational technology, collaborative learning, educational ecosystems, ubiquitous learning, mind maps.

## 1 Introduction

The contemporary world is characterized by the effects of scientific-technological transformations, and undoubtedly is the social, political, cultural and educational impact [2].

In the educational industry, the interactions that emerge by "learning in technological environments" in educational settings have increased considerably; Not only the demand for traditional means of education, now with the advancement of science and technology, is perceived a factor that is provoking educational revolution in the current society of learning called by Stiglitz; Towards the different learning

systems [3].

In this context, the information resource offers the teacher, not only consult the thematic content, but interact with environments that test their knowledge and implement guidelines that mark the theory [4]. Consequently, online learning technologies converge on methodologies, techniques, usage and tools with modern and interactive approaches supported as development technology [5]. Such is the case of b-learning mediated in learning objects in the periodontics theme with interface based on mental maps.

Therefore; The study analysis, which takes into account the characteristics in the b- learning mode of a learning object (case study, activities, resources with evaluation criteria), multimedia software called Hardware Flash Pro (HFP).

The aspect to be considered is the interface applying mental maps, according to the investigation of Tony Buzan (2006), facilitate the visualization and retention of information developing the creativity of the student [6], which makes of HTF a simple, practical and effective software. The research was developed in postgraduate students of periodontics with interdisciplinary management of software engineering.

## 2 Materials and Methods

The development of the present work; The scientific method of Mario Bunge [7] was applied. The design and development of software was used the learning model based on concept maps in learning objects based on a software development model cascade [8]. The development of the learning object is necessary to make the use case. With the purpose of identifying the roles of the actor, ie the activities of those who will participate. In identifying the interactions and actions of the actors, the O.A use case diagram was made, in order to identify the different interactions and actions of actors. The above based on unified UMD modeling language.

**Table 1.** Case Diagram of O.A. use. HFP based UML.

Student	Teacher
Initial navigation	Evaluation of activities
Object analysis	Final evaluation of self-evaluation.
commended activities	
mpletion of activities	
Self-appraisal	

The software was made in Macromedia Flash, which provides tools in multimedia that makes it attractive [9].

**Table 2.** Evaluation of the general MACODA process.

Requirement level
Analysis level
Design level
Implementation level
Evaluation level
Feed back

### 3 Development

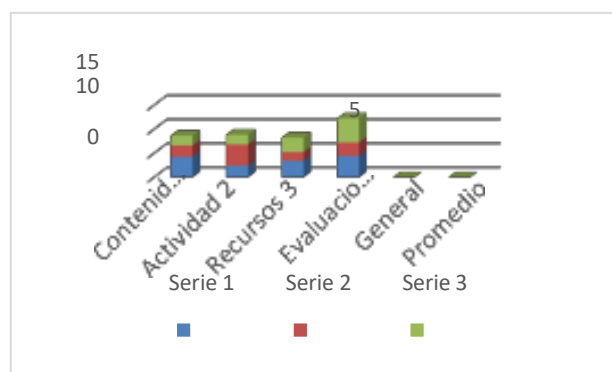
Initially different learning objects are formed under the same established scheme, these objects later formed the granularity; That is to say the union and homogenization of each object, thus arriving at Hardware Flash Pro. HFP combines the characteristics of multimedia and the advantages of mental maps; The interface is managed based on images alluding to the subject, the corresponding text and background music according to the subject of interest. The software contains animations, background music and allows access to the main menu or exit the application.

The interfaces of the mental maps have a practical representation and distribution in such a way that the user becomes familiarized and conceptualized during the analysis in relation to the simplicity of the graphs. The subject matter (analysis of clinical cases, periodontics); Associates the image and the description of the clinical case referenced.

### 4 Results and Discussion

The present research is the product of a practical theoretical foundation generated in learning environments made with experts (interdisciplinary management) at the Cooperative University, Bogotá and Unicomfauca, generating an exchange of knowledge applied in the experimental phase. The type of research is empirical analytic [10], with a sample [11] of 25 students the total of the second cohort in a two-year period.

The data collection instrument was carried out in a survey consisting of 20 questions, considering the evaluation in the subjects (subject, activity, content and resources) followed by questions with response on the Likert scale and entered into the SPSS version 12 system



**Fig. 1.** Graph evaluation result for 25 students (own source).

The variables evaluated were: retention time and learning association were associated with independent variables such as: retention time, content and activities associated with learning, resources and evaluation.

The results obtained are: 80% on a 100% scale, ie greater than 60% of the highest value assigned on the Likert scale; It was concluded that the learning object represents an effective support tool in the teaching-learning process

## 5 Future Work

As a future work, the points with the lowest evaluation are improved: activities and evaluation. In addition, a final test with a larger master size of 30 students in 3 cohorts and correlating 2 or more specialties is carried out with the purpose of strengthening the results and checking Its functionality in different disciplines.

## 6 Conclusions

Technological mediations in virtual learning environments emerge in the construction and deconstruction of different bets in contemporary society such as network society (Castell), society of the spectacle (Debor), society of control (Foucault), society of learning (Stiglitz). The new forms of symbolic and material production transform the subject into his subjectivities; (Donna Haravan) in the so-called "subject cybor" to understand this phenomenon is an academic challenge for those who belong in the current academic community

Mind maps facilitate the visualization and understanding of elements in the interface allowing the student to become familiar with the elements.

The hybrid models of learning in virtual environments and the interface based on mental maps allows to generate comparison between the educational models in education superior according to the fruit of the learning ubiquitous.

## References

1. Begoña, K.: The future of ubiquitous learning. London, Springer, p. 37 (2016)
2. Hernández, S.: Procesos educativos e investigación en la virtualidad. Revista de innovación educativa. Apertura, vol. 4, p. 63 (2016)
3. Margain, L.: Aprendizaje colaborativos en la gestión del aprendizaje. Revista Iberoamericana de educación a distancia, vol. 8, p. 5 (2015)
4. Cruz, G.: Exploración del aprendizaje en los estudiantes haciendo uso de ambientes colaborativos. inteligencia artificial. Revista Iberoamericana de tecnología educativa 8(4), 17 (2014)
5. Milth, J.: Learning Technology Standars Committe. IEEE Standard for Learning Objetivo Metadata. IEEE Standard, p. 63 (2012)
6. Buzan, T.: Mapas mentales. Barcelona, Urano, pp. 206 (2006)
7. Bunge, M.: La ciencia, su método y su filosofía. España, Yauros, p. 68 (1996)
8. Margain, L.A.M.J.: Modelos de aprendizaje basados en mapas conceptuales. UAA, México, UPA, p. 142 (2015)
9. Lindhaert, J.: Introduction to Macromedia Flash. State University Faculty and Staff (2015)
10. Habermas, J.: Conocimiento e interés. p. 38 (1996)
11. Collazos, C.F.: Metodología de investigación. Bogotá. Mc Graw Hill (2014)

# Cryptoanalysis of the 340-bit RSA Algorithm using SBC

Nelson Darío Pantoja<sup>1</sup>, Anderson Felipe Jiménez<sup>2</sup>,  
Siler Amador Donado<sup>3</sup>, Katerine Márceles<sup>4</sup>

<sup>1,2,4</sup> Institución Universitaria Colegio Mayor del Cauca, Popayán, Colombia

<sup>3</sup> Universidad del Cauca, Popayán, Colombia

<sup>1</sup> dariopantoja@unimayor.edu.co, <sup>2</sup> afjimenez@unimayor.edu.co

<sup>3</sup> samador@unicauca.edu.co, <sup>4</sup> kmarceles@unimayor.edu.co

**Abstract.** This work used computer devices selected from requirements mainly aimed at their physical processing components to establish through metrics the importance of the hardware when conducting performance tests on the different cryptanalysis techniques applicable to be RSA algorithm. In order to select the adequate technique for cryptanalysis, performance tests were carried out among those chosen devices mentioned in the adequate technique was selected in terms of time and efficiency. To avoid compromising the integrity of the results, the tests were run similar environments and hardware and software levels, using Kali Linux as operating system and the Python language for the cryptanalysis technique, given compatibility with the three machines and their performance.

**Keywords:** RSA, cryptanalysis, SBCs, devices, server.

## 1 Theory of the Domain and Prior Works

The RSA algorithm is currently one of the most secure [1] to establish communication between an emitter and they receptor, hence, a rupture can lead to many consequences [2]. Its security is due to it being a cryptographic system that uses two keys, one private and one public, which in turn use extremely large primary numbers, generally 2048 bits – as recommended, requiring high computational efforts to decipher it by using factorization methods. With the frenzied progress in technology, multiple devices are available with computationally diverse capacities, which can be used to carry out different tasks and in this case specifically answer the question: what hardware components are important when executing a cryptanalysis process to the RSA algorithm? If cryptanalysis depends on the confrontational resources offered by the machine, it may be stated that greater resources would mean greater efficacy in executing cryptanalysis.

Given that the analysis of the paper Cryptanalysis of RSA: A Survey [3] proposes that the security of the RSA cryptographic system cannot be doubted because to date a devastating attack has been found in the failures that could have occurred are commonly due to poor implementations of the system.

Recently, the task of conducting cryptanalysis to the RSA algorithm has led researchers to using less conventional methods for this purpose; one of the best known

key extraction through acoustics, explained in the paper RSA Key Extraction via

Low-Bandwidth Acoustic Cryptanalysis [4], which although effective in some cases still seems a rather impractical method not applicable to all possible scenarios.

The article Twenty Years of Attacks on The RSA Cryptosystem, [5] BONEH, Dan, lists the most common attacks to accomplish breaking the keys of the RSA algorithm and groups them into four categories, thus, providing cryptanalysis methods candidates for research in this branch.

Mathematical PhD, Hugo Scolnik, in his article Mathematical foundations of the RSA method [6] claims not needing quantum computing to break the keys of the RSA algorithm with a high number of bits, besides providing results measured over time of techniques to decipher said algorithm.

The paper Factorization of 768-Bit RSA Modulus [7] explains the development of the process to achieve the objective of breaking the biggest number of RSA bits known to date; in addition, it mentions the goals previously reached.

### 1.1 Msieve: Factoring Tecnology

Msieve is a complete package of factorization, which automates the mathematical process in addition to choosing the appropriate algorithm according to the size of the number to factorize, being by default the quadratic sieve technique used for the larger numbers, this last technique optimizes the implementation of the choice of the polynomial that relates to the sieve of the numerical field algorithm, this software works on linux-based systems, which is ideal for implementation in SBC's.

## 2 Prior Research

### 2.1 Selection of Devices

To select the devices, a series of activities were undertaken that permitted knowing and classifying – from a set of criteria – the most common low-cost devices. The device selection process established parameters according to the processing capacity related to its CPU power measured in GHz, principally because these components are in charge of performing the arithmetic logic calculations. As a result, the Raspberry Pi 3 was chosen as the optimal, which was most easily accessed for the evaluation, as shown in Figures 1 and 2.

**Table 1.** List of devices used most often in the market.

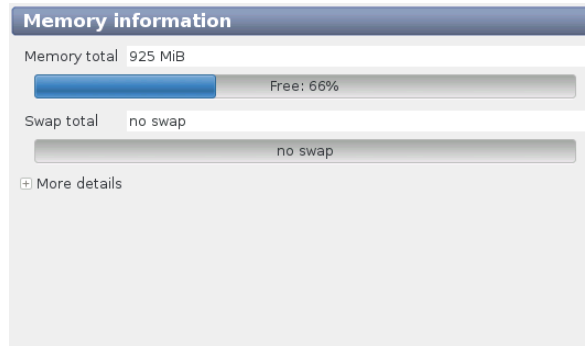
NOMBRE	SOC	CPU	GPU	RAM	OS
<b>Raspberry Pi Model B</b>	Broadcom BCM2835	700 MHz ARM1176JZF-S core (ARM11 family)	700 MHz ARM1176JZF-S core (ARM11 family)	256 MB (shared with GPU)	Debian GNU/Linux, Fedora, Arch Linux ARM
<b>MK802</b>	Allwinner A10	1.5GHz? Cortex-A8	MALI400MP OpenGL ES 2.0	1GB / 512MB DDR3	Android 4.0,Puppy Linux, Ubuntu
<b>Mele A1000</b>	Allwinner	1GHz+ Cortex-A8	MALI400MP	512MB	Android 2.3,ubuntu, Debian,

NOMBRE	SOC	CPU	GPU	RAM	OS
	A10		OpenGL ES 2.0	DDR3	puppy, android ics
Rhombus-Tech A10 EOMA-68	Allwinner A10	1.2ghz Cortex A8 ARM Core	MALI400MP OpenGL ES 2.0 GPU	1gb	-
Gooseberry board	A10	1 Ghz - 1.5 Ghz 1.2 Ghz highest stable	Mali 400 MHz	512MB	Android 2.3, ubuntu, Debian, puppy, android ics
Pineriver H24/MiniX S	A10	1.2GHz	Mali400	512MB	Android 2.3/4.0, Puppy Linux, Ubuntu
Smallart UHOST	Allwinner A10	1GHz Cortex-A8	Mali400	1 GB	android 4.0, Puppy Linux, Ubuntu
A13-OLinuXino	Allwinner A13	1GHz A13 Cortex A8	Mali400	512 MB	Linux
VIA APC	VIA WonderMedia 8750	800 MHz ARM11	OpenGL ES 2.0	DDR3 512MB	Android 2.3
BeagleBoard Rev. C4	TI OMAP3530	720 MHz ARM Cortex-A8	PowerVR SGX	256 MB	Android, Ubuntu, Fedora, ArchLinux, Gentoo
BeagleBoard-xM	TI DM3730	1 GHz Cortex-A8	PowerVR SGX	512 MB	Angstrom, Android, Ubuntu, Fedora, ArchLinux, Gentoo
BeagleBone	TI AM3359 Sitara	500MHZ-USB Powered 720MHZ-DC Powered Cortex A8	SGX530	256MB DDR2 (128MB Optional)	Angstrom, Debian, Ubuntu, Fedora, ArchLinux, Gentoo, Sabayon
PandaBoard	OMAP4430	1GHz Dual-core ARM Cortex-A9 MPCore	PowerVR SGX540	1GB DDR2	Ubuntu, Ångström, Android,Supported by Linaro
PandaBoard ES	OMAP4460	1.2 GHz Dual-core ARM Cortex-A9 MPCore	PowerVR SGX540	1GB DDR2	Ubuntu, Ångström, Android,Supported by Linaro
Cotton Candy	Samsung Exynos 4210	1.2 GHz dual-core ARM Cortex-A9	quad-core 200 MHz Mali-400 MP	1 GB	Android, Ubuntu
CuBox	Marvell Armada 510 (88AP510)	800 MHz ARMv7	Vivante GC600	1 GB DDR3-800MHz	Ubuntu 10.04, Android 2.2, Linux kernel 2.6.x or later Android 2.2.x and later
Hawkboard	TI OMAP-L138	300-MHz ARM926EJ	-	128 MByte	Ubuntu, Fedora, Impactlinux
IGEP v2	Ti DM3730	ARM Cortex A8 1GHz	SGX530 @ 200 MHz	512 M	Android, Angstrom, Ubuntu
IGEP COM Proton	DM3730 (optional OMAP3530)	1GHZ ARM CORTEX A8 (720Mhz for OMAP3530)	SGX 530 (200Mhz) (110Mhz for OMAP3530)	512 MBytes	Android, Angstrom, Ubuntu
IGEP COM Module	DM3730 (optional OMAP3530) A	1GHZ ARM CORTEX A8 (720Mhz for OMAP3530)	SGX 530 (200Mhz) (110Mhz for OMAP3530)	512 MBytes	Android, Angstrom, Ubuntu, ..
Gumstix Overo series	AM3703, DM3730, OMAP3503, OMAP3530	ARM Cortex-A8 Up to 1GHz	-	512MB or 256MB	Ubuntu, Android, ..
Origen Board	Exynos4210	1.2GHz Dual Core Cortex-A9	Mali400 MP4	DDR3 1GB	Android, Ubuntu, ..
Nimbus	Marvel Kirkwood 6281	1.2 GHz	-	512MB	Debian
Stratus	Marvel Kirkwood 6281	1.2 GHz	-	512MB	Debian

NOMBRE	SOC	CPU	GPU	RAM	OS
SheevaPlug dev kit (Basic)	Marvel Kirkwood 6281	1.2 GHz ARM9E	-	512 MB	Ubuntu/Debian
GuruPlug Standard	Marvel Kirkwood 6281	1.2 GHz	-	512 MB	Ubuntu/Debian
GuruPlug Display	Marvell ARMADA 168	800MHz	-	512MB	Ubuntu/Debian
DreamPlug	Marvel Kirkwood 6281	1.2 GHz	-	512MB	Ubuntu/Debian
D2Plug	Marvell PXA510	800MHz	-	1GB	Ubuntu/Debian
Trim-Slice series	NVIDIA Tegra 2	1 GHz	-	1 GB DDR2-667	Android, Ubuntu, ..
Snowball	STEricsson Nova A9500	1GHz Dual Cortex A9	Mali 400	1GByte	Linaro (Ubuntu, Android)
i.MX53 Quick Start Board	Freescale i.MX535	1GHz	-	1GB of DDR3	Linaro (Ubuntu, Android)
Genesi Efika MX Smarttop	Freescale i.MX515	ARM Cortex-A8 800MHz	-	512MB	Ubuntu
FriendlyARM Mini 210s	Samsung S5PV210	1 GHz Cortex-A8	PowerVR SGX540	512 MB	Linux-2.6.35, Android 2.3, 4.0, WindowsCE 6.0
Embest DevKit8600	TI's Sitara AM3359	720MHz ARM Cortex-A8	SGX530	512MBytes	Linux 3.1.0, Android 2.3 and WinCE 7
Embest SBC8018	TI AM1808	375MHz ARM926EJ-S	128MByte	128MByte	Linux2.6.33 and WinCE 6.0
Embest SBC8530	TI DM3730	1GHz ARM Cortex-A8	-	512MByte	Linux2.6.32, Android 2.2 and WinCE 6.0
Embest DevKit8500D	DM3730	1GHz ARM Cortex-A8	-	512MB	Linux2.6.32, Android 2.2 and WinCE 6.0.15
TechNexion Infernopak	TI OMAP3530	600Mhz	POWERVR SGX 530	128 MB	Linux 2.6.x, Windows CE 6.0 BSP or Android
TechNexion Thunderpack	TI OMAP3530	600Mhz	POWERVR SGX 530	256 MB	Linux 2.6.x, Windows CE 6.0 BSP or Android
VIA ARTiGO A1200 Fanless	Chipset VIA VX900	1.0GHz x86 VIA Eden X2 L2 Cache 2MB	VIA Chrome 9	2Gb Up to 4GB DDR3	"ordinary" X86 OS
VIA ARTiGO A1150	Chipset VIA VX900H	1.0GHz VIA Eden X2	VIA Chrome 9	2Gb Up to 4GB DDR3	"ordinary" X86 OS
DMP - eBox 3350MX	-	1Ghz Vortex86MX (i586, no CMOV)	-	512MB	i586 compatible OS
DMP - eBox 3310MX-AP	-	933MHz Vortex86MX+	-	1GB DDR2	i586 compatible OS

CPU information	
Vendor	unknown
CPU(s)	4
Model name	ARMv7 Processor rev 4 (v7l)
Frequency	unknown
L2 cache	unknown
More details	

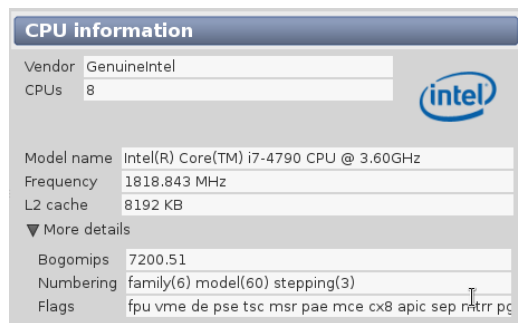
**Fig. 1.** Characteristics of the Raspberry Pi 3 processor (1.2 GHz, according to the Raspberry official web page). Source: Author's information.



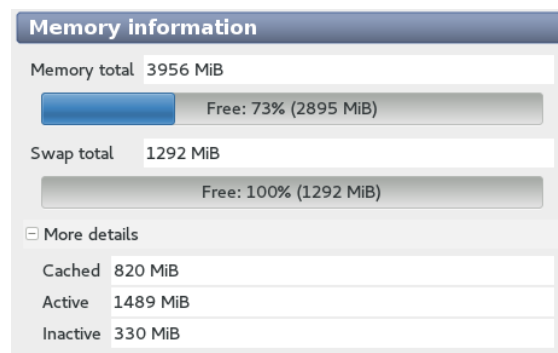
**Fig. 2.** Raspberry Pi 3 RAM memory information. Source: Author's information.

According to the previous criterion, and to compare results, the same test was evaluated on two more devices with superior characteristics over the first.

The second device was a personal computer (PC) with intermediate characteristics in its hardware, as seen in Figures 3, 4 and 5.

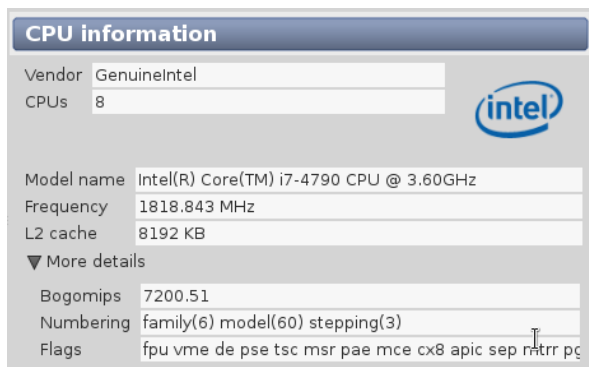


**Fig. 3.** Characteristics of the processor in the intermediate PC. Source: Author's information.

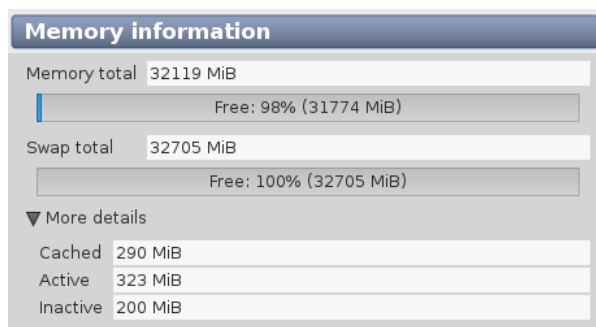


**Fig. 4.** RAM memory information of the intermediate PC. Source: Author's information.

Finally, a computer with features really higher than the previous ones, being this one generally used like server:



**Fig. 5.** Characteristics of the processor in the server. Source: Author’s information.



**Fig. 6.** Server’s RAM memory information. Source: Author’s information.

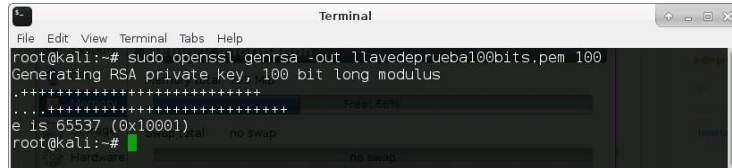
## 2.2 Software

To keep from compromising the integrity of the test results, these were conducted in equal environments regarding software; the three devices used 64-bit Kali Linux 2016.2 as operating system. To execute the cryptanalysis, the msieve open code program was used, which is developed under C language and can be executed in multiple operating systems. With regard to the generation of keys, such was carried out with OpenSSL, which – in turn – generally distributes cryptographic options to web sites for secure HTTPS access.

## 3 Experiments and Results

The first step established the controlled work environment, which was a computer laboratory with the physical space to conduct the corresponding experiments. This work setting has the selected computer equipment already mentioned. The experiment conducted

were tests with keys generated with OpenSSL; the keys generated were of 100, 256, and 340 bits (Fig. 7).



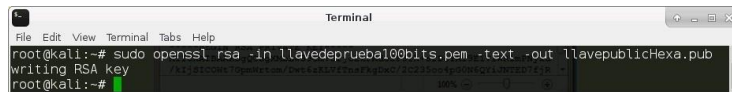
```

root@kali:~# sudo openssl genrsa -out llavedeprueba100bits.pem 100
Generating RSA private key, 100 bit long modulus
.....
e is 65537 (0x10001)
root@kali:~#

```

**Fig. 7.** Example of key generated with OpenSSL. Source: Author's information.

To convert the keys generated into a hexadecimal format, the following OpenSSL commands were employed (Figs. 8, 9, 10, and 11).



```

root@kali:~# sudo openssl rsa -in llavedeprueba100bits.pem -text -out llavepublicHexa.pub
writing RSA key
root@kali:~#

```

**Fig. 8.** Example of a hexadecimal key. Source: Author's information.

```

Private-Key: (100 bit)
modulus:
  0afc48b5385f57a673422fba49 =>870359746064855796858720598601
publicExponent: 65537 (0x10001)
privateExponent:
  09:19:b3:d5:c4:4c:59:dc:bd:74:b2:1e:81
prime1: 3021521809 (0xb418c391)
prime2: 4212375353 (0xfb13bf39)
exponent1: 1633169041 (0x61582e91)
exponent2: 915059401 (0x368ab2c9)
coefficient: 602854641 (0x23eed4f1)
-----BEGIN PUBLIC KEY-----
MCgwDQYJKoZIhvcNAQEBBQADFWAwFAINCvxIThFV6ZzQi+6SQIDAQAB
-----END PUBLIC KEY-----

```

**Fig. 9.** 100-bit hexadecimal key generated with Open SSL. Source: Author's information.

```

Private-Key: (256 bit)
modulus:
  00b2b980c4233da33323fb7f7a95040d8229593b8f7580897f133e09993b8ea091
  =>8083944246619864058992914082385909967586109685764138081385009245529733767313
publicExponent: 65537 (0x10001)
privateExponent:
  00:8c:eb:fd:df:29:96:61:47:62:b8:dc:84:70:59:
  38:b8:35:b6:ad:f2:29:ef:83:a8:2c:1e:15:12:d8:
  e3:20:01
prime1:
  00:e7:8c:58:db:51:04:31:46:cf:ee:2f:23:64:86:
  4a:91
prime2:
  00:c5:99:20:9f:ff:fe:f0:a8:39:b0:9e:a2:bb:d1:
  f6:01
exponent1:
  09:7e:50:9a:55:5d:05:a4:30:9c:44:64:80:17:9d:
  71
exponent2:
  57:02:fe:1d:d6:c1:b1:c1:b2:5d:b7:0d:5b:fd:b2:
  01
coefficient:
  08:6b:50:11:5e:23:1d:c5:47:54:a3:95:7e:6d:37:
  dd
-----BEGIN PUBLIC KEY-----
MDwwDQYJKoZIhvcNAQEBBQADKwAwKAIhALK5gMQjPaMzI/t/epUEDYIpWtuPdYCJ
fxM+Czk7jqCRAGMBAE=
-----END PUBLIC KEY-----

```

**Fig. 10.** 256-bit hexadecimal key generated with Open SSL. Source: Author's information.

```

Private-Key: (340 bit)
modulus:
0b23057297dab86cd4a965e01854d7f26bed01adf289d2b34aed7ed8f636a8d5ce23d6b33c4a0f6a69f9b3
->1558974590155323540076393624175863694369612223119291374255782520022941558818389212336819091060799699379

publicExponent: 65537 (0x10001)
privateExponent:
01:85:42:95:26:ce:a2:27:99:d1:97:2b:45:a7:2f:
e4:d5:7f:82:9f:31:61:5d:68:46:0b:2f:c9:fb:58:
dd:85:92:93:3b:8c:a8:09:fe:3e:7b:b5:c1
prime1:
03:b4:2d:48:9b:f8:53:33:a2:5d:73:97:0c:71:c9:
92:f0:96:0a:e6:f2:4b
prime2:
03:01:c0:9e:15:d9:af:ab:61:92:b9:4c:ed:1a:fd:
03:73:9d:85:ad:b5:39
exponent1:
01:bb:77:a2:8a:30:6e:d9:ab:8b:01:d1:17:e4:f0:
5e:65:00:07:e1:54:59
exponent2:
00:ff:8a:ef:b0:77:55:4f:83:14:0f:ba:4f:18:df:
98:4e:c0:a3:c9:78:59
coefficient:
01:70:38:c5:ba:ef:4c:bb:de:fb:d2:f9:a2:a2:ac:
09:87:41:56:d6:71:63
-----BEGIN PUBLIC KEY-----
MEYwDQYJKoZIhvcNAQEBBQADAwgIrCymFcpfauGzUqXkgGFTX8mvtAa3yidKz
Sul+2PY2qXOI9azPEoPam5swIDAQAB
-----END PUBLIC KEY-----

```

**Fig. 11.** 340-bit hexadecimal key generated with Open SSL. Source: Author's information

By using the integer factorization program with the Number Field Sieve (NFS) factorization algorithm denominated msieve, proceed to enter module  $n$  from the key of public knowledge, which was subjected to the factorization attack to obtain the key with which to encrypt the message that will be transmitted (Fig. 12.)

```

Msieve v. 1.53 (SVN Unversioned directory)
random seeds: a79b8e6d 91c6df8e
factoring 1558974590155323540076393624175863694369612223119291374255782520022941558818389212336819091060799699379 (103 digits)
no P-1/P+1/ECM available, skipping
commencing quadratic sieve (103-digit input)
.
.
.
p52 factor: 1124994410756636672075595149516955324427504416830777
p52 factor: 1385762076014942273883725154208627052395961971569227
elapsed time 05:14:26

```

**Fig. 12.** Example of factorized key with msieve. Source: Author's information.

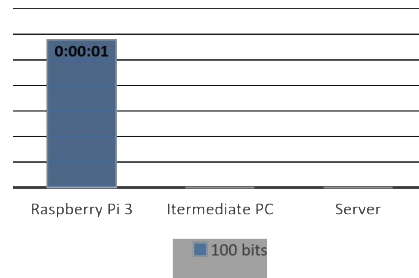
The msieve program registers its activity and results in a file called msieve.log, where it is evident that effectively module  $n$  was decomposed into two prime factors assumed as  $p$  and  $q$  and which can be used to obtain the key with which the message is encrypted.

## 4 Conclusion and Future Work

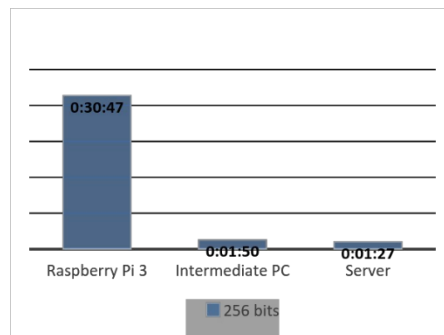
From the experiments conducted, it may be concluded in the first place that the architecture of a device's processor is quite relevant when carrying out a cryptanalysis process on the RSA algorithm through factorization, given that it influences directly on the time required for this operation but is not an impediment to be carried out. Upon evidencing the device's relevance regarding time, a comparative graph was made showing the degree of performance (Figs. 13, 14, and 15).

It is evident that the NFS encryption algorithm used through msieve behaves much more efficiently as the frequency of the processor increases.

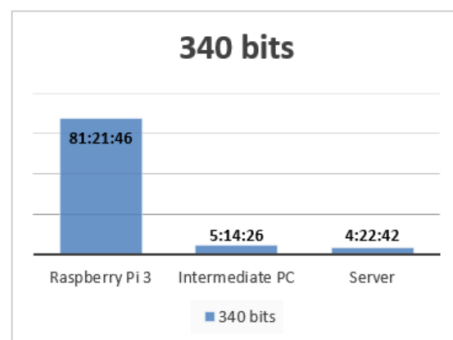
We can also observe the behavior of the RAM memory, which eventually is different in the three devices; in addition, relatively little use is noted in relation to the total memory (Fig 16).



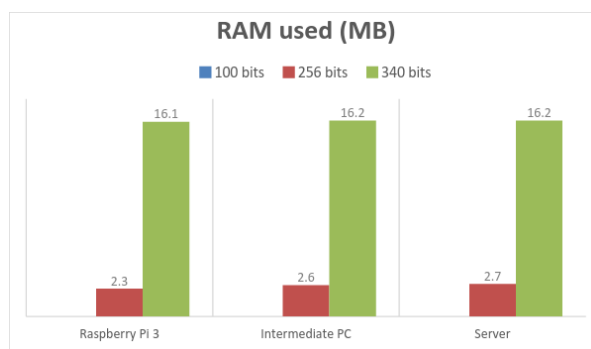
**Fig. 13.** Graph of Devices vs. Time for 100 bits. Source: Author's information.



**Fig. 14.** Graph of Devices vs. Time for 256 bits. Source: Author's information.



**Fig. 15.** Graph of Devices vs. Time for 340 bits. Source: Author's information.



**Fig. 16.** RAM memory used by the device. Source: Author's information.

Based on the results, we want to provide, in a basic way, the recommendation of the use of SBC devices to perform cryptanalysis to the RSA algorithm, given that they can perform, although less efficiently, mathematical operations with the advantage of their cost in the market.

**Future work:** Evaluate the behavior of cryptanalysis on the RSA algorithm in a distributed system, where the selected attack is divided by the different nodes that make up the system. Establish a theoretical time according to the numerical complexity of the technique selected and the hardware characteristics of the device chosen to perform the cryptanalysis to, thus, predict the time of the cryptanalysis used with the different RSA key sizes. Compare cryptanalysis times on the RSA algorithm by using CPU and GPU with keys of at least 340 bits.

**Acknowledgments:** to the Cryptography group and to the GTI Research group of the University of Cauca and to Beta Bit seed of the Research and Development group in Information Technology of the University Institution Colegio Mayor del Cauca, for the support provided for the development of the project.

## References

- 1 Zhou, X., Tang, X.: Research and implementation of RSA algorithm for encryption and decryption. In: Proceedings of 2011 6<sup>th</sup> International Forum on Strategic Technology, vol. 2, pp. 1118–1121 (2011)
- 2 Rivest, R.: A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Mag. Commun. ACM* 21(2), pp. 120–126 (1978)
- 3 Cid, C.: Cryptanalysis of RSA: A Survey. SANS Inst. InfoSec Read. Room (2003)
- 4 Genkin, D., Shamir, A., Tromer, E.: RSA Key Extraction via Low-Bandwidth Acoustic Cryptanalysis. In: Advances in Cryptology - CRYPTO 2014: 34<sup>th</sup> Annual Cryptology Conference, Santa Barbara, CA, USA, August 17–21, 2014, Proceedings, Part I, J. A. Garay and R. Gennaro, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 444–461 (2014)
- 5 D. Boneh, Twenty Years of Attacks on the RSA Cryptosystem. *Not. Am. Math. Soc.*, vol. 46, pp. 203–213 (1999)
- 6 Scolnik, H.: Fundamentos matemáticos del método RSA. (2004)
- 7 Kleinjung, T.: Factorization of a 768-Bit RSA Modulus. In: Advances in Cryptology- CRYPTO 2010: 30<sup>th</sup> Annual Cryptology Conference, Santa Barbara, CA, USA, August

- 15-19, 2010. Proceedings, T. Rabin, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 333–350 (2010)
- 8 Alimoradi, R., Arkian, H.: Integer Factorization Implementation. ICTACT Journal on Communication Technology 7(2), pp. 1310-1314 (2016)



# Security Validation with Owasp Mobile for the Data Protection in Oral Health

Katerine Márceles<sup>1</sup>, Clara L. Burbano<sup>2</sup>, Gustavo Uribe<sup>3</sup>, Diana Burbano<sup>4</sup>

<sup>1,2,3</sup>Grupo TIC-Unicomfacaucá, Corporación Universitaria de Comfacaucá, Popayán, Colombia

<sup>4</sup>Universidad Cooperativa de Colombia, Popayán, Colombia

{<sup>1</sup>kmarceles, <sup>2</sup>cburbano, <sup>3</sup>guribe}@unicomfacaucá.edu.co

<sup>4</sup>caritoburg@yahoo.com

**Abstract.** The paper presents an analysis of requirements of information security in mobile applications focused on oral health. We studied the security of a mobile application developed in Android, referring to national and international standards, legislation and the Owasp Mobile recommendations. The current increasing use of mobile devices and the ubiquity of information through Internet allows the unauthorized access to information. Also, oral health mobile applications face this problem. Therefore, they usually don't include private information required in some health processes. This paper faces the problem of the information reliability and security in health mobile applications, performing security tests aligned with the risks identified for the Owasp mobile.

**Keywords:** security, data protection, mobile applications, mHealth, legislation, OWASP.

## 1 Introduction

Guaranteeing security of mobile applications is an imperative due to diverse attacks on their integrity, confidentiality, and availability. The current problem with mobile applications is lack of protection and not treating information according to standards and legislation of the country of origin. In order to mitigate this problem, good practices of mobile application programming have been developed, documented in the Open Web

Application Security Project (OWASP) Top Ten of Most Critical Web Application Security Risks [1]. It addresses problems like information leakage, missing or broken authentication of the application user, enabling phishing, attacks through cross-site scripting and/or SQL injection, among others. The technological development presented in this article guides to generate an application that at least complies with the legal data protection requirements of the country of origin and is built under safe development methodologies aligned with OWASP for mobile phones. The country of origin for this study is Colombia. However, the principles followed can be applied to other countries as well. The current study was performed on a mobile application named OralHealth with the function of calculate the O'Leary index based on the stored information. Therefore, appropriate measures to protect users' security and privacy are inevitable.

## 2 Mobile Applications in the Field of Oral Health

In recent years, the term "mobile health" [2] has been used for referring to the use of mobile devices in the health care. Currently, there is a variety of apps (mobile applications) in the area of oral health, where users are dentists, but also children and adults, who intend to take care of their teeth and to have fun while deploying their mobile. A list of apps is shown in Table 1.

**Table. 1.** List of mobile applications for the Oral Health [3,4].

Application	Description
Brush Dj	Application offering a play-list during two minutes while tooth brushing is performed
Brush	This application teaches the correct tooth brushing trough different games
Dental Expert	Application that respond to the frequent asked questions of the users and give advices about topics related with the oral health as: Teeth whitening, oral self-care and emergencies.
Tooth protection tips	This application offers tips for keeping a perfect smile
BrushyTime	Application for children teaching the tooth brush trough draws, games and a digital clock
Dental Care Aid	Application that describe procedures related with the oral health using illustrations
Philips Zoom	This app allows the virtual teeth whitening

The Food and Drug Administration [5], define that the software executable in a mobile platform can be used as medical device. In the mHealth App Developer Economics [6], the 3% of the apps are published in the health category in the Google Play, iTunes and Microsoft Phone stores.

Also, is highlighted the presence of smart mobile devices in countries as: Venezuela (20%), Chile (18%), Argentina (17%), Brazil (16%), Mexico (15%), Perú (9%) and Colombia (8%). For the 2013 year, researchers discovered the fact that Android have a 45% of participation in the market [6][7].

According to the 2015 statistics in Colombia, mobile devices were used by 95.6% of the population [8]. It is important to note that around 700,000 apps are available for download [9].

On other hand, 3% of them are apps related to health, which have reached about 44 million downloads a year [10] [11]; In 2018, it is estimated that about 50% of the 3.4 trillion mobile devices will use health apps [12].

### 2.1 Information Security in Mobile Applications

Information security is defined as preservation of confidentiality, integrity, authenticity and availability of information managed by, e.g., mobile applications. It requires technical measures to guarantee authenticity, reliability and non-repudiation of access to hardware (e.g. tablets and/or smart phones) and the software (data entered, stored, processed and communicated by OralHealth application).

When talking about information security on mobiles, it is necessary to identify their vulnerabilities (possible attacks) as well as the risks (probability of occurrence and resulting damages) when using the devices. Possible security risks are the infection of the device by malware, or unauthorized access to, theft or modification of information. Bad practices in the development of the applications increase the aforementioned risks. In order to minimize vulnerabilities and risks it is important to consider the operating system, technologies and communication networks, and to deploy good development practices and related standards [13].

OWASP Mobile Security Model is a tool intended to give developers the resources they need to build and maintain secure mobile applications. Furthermore, it allows the classification of mobile security risks providing necessary development controls to reduce the risk of attacks against the applications [1]. For a secure development of the OralHealth application based on the Android operating system, the OWASP Mobile Top Ten guidelines [14] as well as the NowSecure Chapter 7 Mobile Applications Development Manual for Android systems have been used [15]. Furthermore, the national and international legislation safeguarding data stored by the user was considered. The Android platform was used because of its significant growth in the last years with respect to the other platforms. For example, in 2016 Android could score an increase of use by 10.3% in relation to IOS and others [16].

Additionally, it was required that the OralHealth application complies with the legal guidelines of the Colombian government and international standards. In Colombia, the Online Government Manual requires for 2018 that every application must guarantee the four dimensions making up security and privacy of information measures designed to protect information against unauthorized access, unintended use, unauthorized or illegal disclosure, interruption or unauthorized destruction [7].



**Fig. 1.** The top ten information security risks for mobile apps [1].

The top ten information security risks for mobile apps was published in 2014 by OWASP, based on the statistics of vulnerabilities. The Fig. 1 shown this top ten security

risks. Furthermore, the apps should follow the next steps in order to analyze its level of information security [18]:

1. Compilation of information: The scope of the application is defined.
2. Static analysis: The security in the source code of the application is verified.
3. Dynamic analysis: The security of the application is analyzed in execution over a device or an emulator [19].

## **2.2 Legal and Regulatory Aspects in the Area of Health**

Internationally, the health entities responsible for the proper management of patients' medical information align their processes to regulations and standards such as HIPAA, COBIT 5, ISO 27002, CALDICOTT, ITU-T, and HL7. They provide guidelines for regulating the exchange of personal data. Another relevant regulation is the Organic Law 15/1999 of Spain related to the automated processing of personal data LOPD [20]. Following, the selected standards implemented in OralHealth are briefly described. HIPAA [21] (Health Insurance Portability and Accountability Act), this law defines the policies for protecting the confidentiality, integrity and availability of the patient information. ISO / IEC 27002:2013 defines good information security practices, addressing confidentiality, integrity, authenticity and availability of information as well as the security of information systems involved [22].

Similarly, there is a related national legislation on data protection in place in Colombia to guarantee confidentiality, integrity and availability. Examples are: Law 1273 from 2009, article 269F, on violation of personal data by entities who, without being entitled to do so, for their own benefit or for a third party, obtain, compile, subtract, offer, sell, exchange, send, purchase, intercept, disclose, modify or use personal codes, personal data contained in files and/or databases [23].

In Article 10 of Decree 1377 from 2013, which regulates Law 1581 from 2012, the National Health Institute, as handler of personal data obtained through a website and / or any other type of device, requests its users' authorization when deploying personal data in accordance with the privacy policies that have been established under the terms of Law 1581 from 2012 on Protection of Personal Data in Colombia. It should be noted that personal data provided may be processed, collected, stored, used, deleted and / or updated in accordance with terms and conditions of the privacy policies established by the National Health Institute [24].

On the other hand, the congress of the Republic of Colombia promulgates in the 599 law of July 24 of 2000 [25] the crimes against individual liberty and other guarantees, including the violation of the privacy, reservation and interception of communications. The violation of the privacy in the Colombian communication enforce the data protection declared in the 192 articles.

## **3 Methods**

The work realized is the type descriptive and experimental performing tests of dynamic analysis for identify the vulnerabilities of the mobile applications based on the OWASP mobile recommendations.

Taking into account the different aforementioned concepts, we continued verifying for the designed OralHealth application the compliance with Colombian minimum

legal and regulatory requirements and international standards. Furthermore, we ensured that the development process followed good practice in safeguarding personal data. The application was verified using the checklist structured under article 269f of Law 1273 from 2009 and Decree 1377 from 2013 of Law 1571 from 2012 about the policies of use, i.e. terms and conditions the application follows at the time of registration. Furthermore, computer security software tools such as zAnti, YSO Mobile Security Framework, ZAP OWASP and Wireshark have been applied.

Those tools enabled the performance of intrusion tests, allowing the verification of some controls of ISO / IEC 27002: 2013 related to Clause 9 Access Control, Clause 10 Cryptography, Clause 14 Acquisition, Development and Maintenance of Information Systems, and Clause 18 Compliance with Regard to the Legal Part. It should be noted that the YSO Mobile Security Framework as a hybrid tool generated several false positives and negatives which were manually verified to corroborate them. The Top Ten of the OWASP Mobile Security Project [9] have been checked for each of the 10 risks identified in the 2014 list: *Weakness in the server-side controls of the application*, *Insufficient storage in the transport layer*, *Insufficient transport layer protection*, *Unintentional data leak*, *Poor authentication and authorization*, *Broken cryptography*, *Client-side injection*, *Security decisions via untrusted entries*, *Handling of inappropriate sessions*, and *Lack of protection of binaries*. The YSO Mobile Security Framework have been complemented by social engineering techniques.

## 4 Results

Verifying the risks determined by OWASP, by widely used information security standards, and by the data protection legislation in Colombia, the following evaluation results were obtained for the OralHealth application:

**M1 Weakness in the server-side controls of the application:** For the exploitation of this vulnerability SQL injections were made, allowing the validation of the corresponding text fields. It is evident that the mobile device connects to the server application remotely, which has sufficient security controls, since the application can manage non-validated or malicious data preventing the SQL injection.

**M2 insufficient storage in the transport layer:** The application stores the data in a secure way, i.e., it does not store them in the device but in the server, and it does not store temporary data. The stored data was encrypted with the AES 256 encryption standard, and the PBKDF2 function allowed the generation of strong keys.

**M3 insufficient protection in the transport layer:** The application in the device is connected to the server, transmitting information over an encrypted connection. The SSLSocketFactory for secure SSL/TLS channels is used to validate the server identity, so reducing the interception risk for the stored data.

**M4 unintentional data leakage:** When updating the operating system, software frameworks and application did not change the behavior. Furthermore, the status of the back doors was checked.

**M5 poor authentication and authorization:** The application provides adequate and necessary levels of authorization and authentication by implementing secure pass-

words of at least 6 digits with alphanumeric and numeric characters. Furthermore, the location of the memory storing the password for hash calculation is deleted.

**M6 cryptographic weakness:** The application performs an adequate encryption of the information stored and transmitted (from or to the device), as evidenced above. Additionally, the executable code is obfuscated by randomization of the design address space (ASLR).

**M7 injection from the client side:** The mobile device application has security controls for data entry and for sending data to the server. It should be noted that this is one of the risks that took a little longer time for assurance, as black and white box tests (Sql injection attacks, Java Script injection (XSS), Fuzzing and inclusion of local files) had to be performed. Those tests allow verifying the integrity, confidentiality and availability of the information. The queries were parameterized and validated to mitigate the risk of Sql and Fuzzing attacks, deactivation for any WebViews of the File System Access and JavaScript.

**M8 security decisions via untrusted inputs:** The application receives only validated data, given the different controls previously implemented.

**M9 inappropriate session management:** The application has adequate security levels, so that the user session cannot be intercepted and / or overridden. This is done by implementing security mechanisms, among them a check at the beginning of each activity.

**M10 Regarding the lack of binary protection:** It could be verified that there were no changes of the binaries of the application and no modification of the behavior of the binaries during the different training tests, including checks regarding SU (superuser) command as well as certificate verification. In this same sense, the checklist designed under Colombian legislation and aligned with the selected international standards was applied, and its compliance was verified, in addition to the safe development related to the controls defined in the ISO/IEC 27002:2013 and the policies of secure development defined by OWASP, which allowed the assurance of the data stored by the user. On the other hand, it was verified that the OralHealth application can become vulnerable through inadequate configuration of social engineering techniques. Therefore, it is suggested to follow good practices and the recommended installation guide.

Finally, the application OralHealth is in line with the guidelines for mobile insurance development, the specific risks by the Owasp, standards of information security and protection legislation of data in Colombia. Many of the security flaws are not due to operative system updates, but non-implementation of secure application development. Dice to the above the mobile application developers for Android should start using methodologies aligned to standards to provide a degree of computer security to users of the platform.

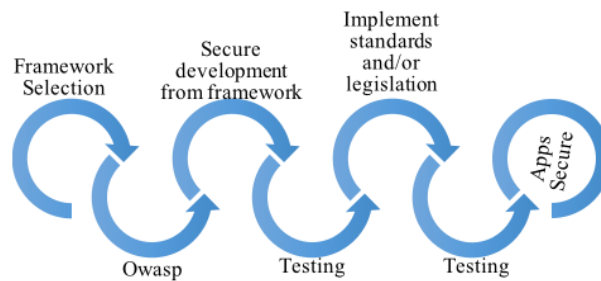
## 5 Discussion

It is important to note that the alteration of information stored in the application can impact the integrity of the user, leading to fraud and / or impersonation through the application. It may involve the alteration and disclosure of personal and clinical data on the user's oral health, e.g.: name, email, login, password, and the O'Leary clinical

records. These are protected by Law 1581 on data protection from 2012, Law 1266 on Habeas Data from 2008 and Law 1273 on cybercrimes from 2009. The legal repercussions of that bring the compulsory of secure development in mobile phones and in any type of application.

## 6 Conclusions

During the process of implementing the Colombian data protection standards and legislation, aligned with the e-Government Manual with its fourth information security dimensions, it was determined that in order to comply with these challenges, it is necessary to implement a secure development process. That process must be parameterized according to the latest identified risks, further mitigating the risk for the stored information. In summary, the model shown in Fig. 2:



**Fig. 2.** Safe agile mobile development.

The model presented in Fig. 2 intends a mobile application with the minimum-security parameters able to safeguard the information in accordance with legal guidelines.

## References

1. Owasp Mobile Security (2016)
2. Fiordelli, M., Diviani, N., Schulz, P.: Mobile Health Research Mapping: A Decade of Evolution. *Journal of Medical Internet Research* (2013)
3. Conozca 5 aplicaciones móviles para cuidar de la salud bucal (2016)
4. Aplicaciones móviles para cuidar tu salud dental (2016)
5. U.S. Food and Drug Administration: CDRH-CBER (2015)
6. Research2guidance: mHealth App Developer Economics (2015)
7. Ramos, S.: Android amplió su dominio a nivel mundial mientras iOS cae. *Social Geek* (2016)
8. Population using mobile devices in Colombia (2015)
9. Administrativo Estadístico Nacional de Estadística: Indicadores básicos de tenencia y uso de tecnologías de la información y comunicación. *Boletín Técnico*, Colombia (2016)
10. Ifrach, B., Johari, R.: Pricing a bestseller: sales and visibility in the marketplace for mobile apps. *ACM SIGMETRICS Performance Evaluation Review* (2014)

11. Fox, R., Cooley, J., McGrath, M., Hauswirth, M.: Mobile health apps - from singular to collaborative. *Stud Health Technol Inform* (2012)
12. Santamaría-Puerto, G., Hernández-Rincón, E., Suárez- Obando, F.: Use of mobile health applications with Internal Medicine patients at the Regional Hospital of Duitama, Boyacá, Colombia. *Cuban Journal of Information in Health Sciences* (2016)
13. Shuren, J.: FDA's role in the development of medical mobile applications. *Clinical Pharmacology & Therapeutics* (2014)
14. Villegas, G.: La seguridad en aplicaciones móviles: estrategias en el mundo actual (2016)
15. OWASP Mobile Security Project: Top 10 mobile risks (2016)
16. nowsecure.com: Android: Secure mobile development (2016)
17. Sergio, R.: Android expanded its dominance globally while iOS drops. <http://socialgeek.co/moviles/android-amplia-dominio-nivel-worldwide-ios-cae> [Last Accessed: 22 11 2016]
18. Ministerio de Tecnologías de la Información y las Comunicaciones: Decreto 2573 de 2014, Colombia (2014)
19. Leonardo-Ramírez, L.: Estrategia de validación para aplicaciones móviles de salud. *IAI, Actas de Ingeniería*, vol. 2, pp. 325–333 (2016)
20. Díaz, S.: Mejores prácticas en las pruebas de aplicaciones móviles, *ATSistemas*, España (2013)
21. Informáticos Europeos Expertos: Protección de datos de carácter personal (2016)
22. Ayuda Legal: Ley Hipaa y la confidencialidad de la información médica (2016)
23. International Standards Organization, ISO/IEC 27002 Information technology – Security techniques – Code of practice for information security controls (2013)
24. Congreso de Colombia: Ley 1273 de 2009 - Legislación de Delitos Informáticos, Bogotá
25. Gobierno de Colombia: Authorization for the Processing of Personal Data National Institute of Health (2016)
26. Secretaría del Senado de Colombia: Ley 599 de 2000

# Conceptualization of Serious Games Used for the Dissemination of the Historical and Cultural Heritage in Colombia

Claudia Sofia Idrobo<sup>1</sup>, Maria Isabel Vidal<sup>2</sup>, Katherine Marcelles<sup>3</sup>,  
Clara Lucia Burbano<sup>4</sup>

<sup>1,2,3</sup> Institution University College most of Cauca, Popayan, Colombia

<sup>4</sup> Foundation University of Cauca, Popayan, Colombia

<sup>1</sup> sofidrobo, <sup>2</sup> mvidal, <sup>3</sup> kmarcelles [unimayor.edu.co], <sup>4</sup> clara\_893@hotmail.com

**Abstract.** Nowadays the use of new technologies of information and communication ICT, together with new paradigms such as the use of serious games, bring new elements in the dissemination of the cultural heritage of cities allowing it be addressed in a way different to those used traditionally, motivated by social and cultural changes for impact to society. This article presents a conceptualization of serious games applied or used in the field of the dissemination of the cultural heritage in Colombia, emphasizing the actors working in the area such as the development of dedicated software companies the creation of serious games whose purpose is to disseminate history and/or Colombian culture and local academic (universities and research groups) providing knowledge in the area, to achieve these objectives we did a search of scientific articles clasificandolos by the themes of the authors creating, an initial draft for the characterization of serious games focused on the dissemination of the cultural heritage, as results review is presented a set of focused serious games to the dissemination of the cultural heritage and a set of features basic operation that will be input to the future work proposed in this article.

**Keywords:** serious games, cultural heritage, edutainment, educational content.

## 1 Introduction

This article is based on a review of theoretical, where the main focus is the exploration of serious games, understood as a product of technology that seeks as well as the entertainment and fun as basic game elements, taking into account aspects as the information on a specific theme, product, or service. Serious games are introduced as a new modality, where the objective is to provide the user with an innovative and fun environment whose added value is, train, learn or even get advertising. According to Mercano B [1], this has allowed this type of technology to be applied in different fields such as: medicine, the military, education, research, advertising among many others. Serious games are classified according to their communicative purpose, oriented theme, target audience and/or work area explored and are focused to communicate, generate valuable information and some awareness about various themes. In this case we have

studied the games that seek to inform and disclose the historical and cultural heritage in Colombia. This type of game is listed under the category of Edutainment, which is understood according to [2] as a combination of methods and types of training that combines the presentation of educational information with elements of entertainment. Its objective is to motivate students in the learning process to increase the capacity of retained information and the ability to transfer to different fields of application.

In recent years, the dissemination of heritage in different countries and especially in Colombia, has taken importance and interest due to the progressive loss in new generations of knowledge of the history and the cultural aspects of the cities, losing an intangible tradition of identity of peoples.

The article is structured as follows: section II presents the conceptualization of serious game, cultural heritage and value of patrimonial serious games; section III describes the methodology that was followed in the process of exploration and finally, section IV presents the results of the review where they are different games and a description of basic characteristics, in which it was in both has developed products in Colombian companies focused on the development of this type of products and research groups with interest in serious games line.

## **2 Serious Games Fundamentals**

This section sets out the general concepts on which is based the research: initially becomes a theoretical approach to the concept of serious play and the relevant elements framing, then a brief description is made of the important aspects of the cultural heritage and as they are regulated in Colombia and finally the third part presents serious games that have been used in the dissemination of the cultural heritage and some features as the platform and the availability of the game.

### **2.1 Serious Games**

Serious games are a technological tool with a purpose that goes beyond having fun, including purposes as educate, train or inform as opposed to a common game, which only seeks to have a good time. The word 'play' is used to represent everything about fun, which in some scenarios could be interpreted away or in opposition to the concept of "seriousness", [3], the game is considered by some authors and sectors of society as a way of enjoy the time free [4] and attributed to him teaching elements that contribute to other fields of action. [5] [6], table 1: classification of types of game, mentioned four existing classifications according to [7] on the game.

The term "serious" indicates the responsibility, good sense, reality and actions included within the product; by joining these two concepts will get the term "Serious games": according to Ludus "these games can be played seriously or casually the interest will have an explicit sense by its educational purpose" [8]. This serious game concept allows users take a roll in from the mechanics and the dynamics that the game has immersed will have the opportunity to experience the success of achieving goals and overcome challenges, and in the case of errors, deal with it as a mechanism for improving and learning [8]. Serious games are used for several purposes, between what is: e-learning, training, simulation, collaboration, advertising, business model. They are

used in the industry and sectors such as: military defense, education, business, scientific exploration, health care, planning cities, engineering, religion, tourism and virtual conferences [9]. In Fig. 1: Serious games according to [10] relationship, the relationship of existing categories of serious games is presented.

**Table 1 .** Overview on the classification of types of game [7].

Classification of Game Typologies	
Psychomotor Games	Body knowledge
	Sensoriomotor
Cognitive Games	Manipulative [construction]
	Exploratory or discovery
	Attention and memory
	Imaginary games
	Language games
Social Games	Symbolic or fictional
	Of rules
	Cooperatives
Affective Games	Role play or dramatic games
	Self-esteem



**Fig. 1.** Relationship of serious games source: TIA-017]

To better understand Fig. 1, it is necessary to understand the terms that make up the great dimension of serious games, initially appears the concept of e-learning, according to [10], refers to the relationship between the education and training via the Internet. This type of teaching online allows the user interaction with the material through the use of various computer tools, concept that is supported through the development and the advancement of information and communication technologies ICT where the information society plays an important role as it seeks an interactive learning. by users (students) who finally are those who receive significant benefits such as the large number of accessible content to accelerate the training process, increasing access to training processes and reducing costs in most of cases; Similarly in the graph shown included the concept of gamification which indicates the application of mechanics and dynamics of the game which aims to motivate people. According to

[11], the gamification is an instrument linked to business and marketing used or applied to positively contribute to recruitment processes in different themes, promotion of products or services, even beyond this promote the its adoption, [10], the gamification and serious games, try to take the best from the games and video games to support different activities in the real world and generate useful applications to users through the dynamics and the game mechanics mainly in marketing and other areas.

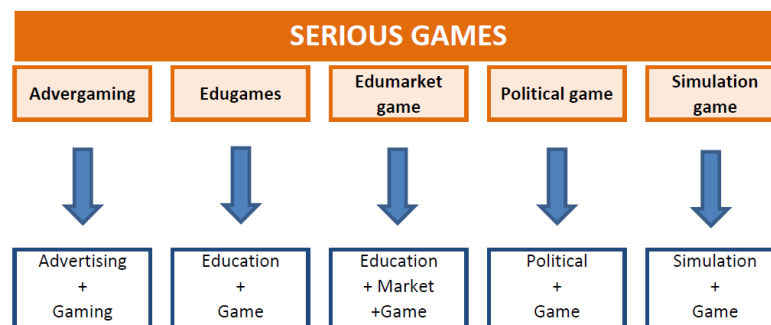
Finally, in Fig. 1: Classification of serious games is summarized as follows by answering a basic response to the above at the beginning of the article: product can be considered as a serious? Serious games are a proposal that combines the previous tools. In a final formula might look like this (see Fig. 2):



**Fig. 2.** Composition of serious games.

The Gamification as motivational strategy seeks to strengthen it and so it has begun to include stories that will positively impact the process and contributing to the improvement of digital content.

The Fig. 3, according to [12], presents the classification of serious games, which will help to understand how located on them focused games at the cultural heritage.



**Fig. 3.** Serious games for cultural heritage.

According to Fig. 3, relates the familiarity of serious games in the category of educational games [13], and this close relationship is where it is considered position, classified at the same time in Edutainment. The concept of "Edutainment" [14] involves the incorporation of "educational resources" in teaching-learning processes to encourage and facilitate the achievement of the goals and in this way make fun and innovative process for who you are carrying out.

## **2.2 Heritage**

In Colombia the entity that regulates everything related to heritage and culture is the Ministry of culture [15], but the regime or special regulation of the Cultural heritage of the nation, is now fully integrated in the 1185 Act of 2008, which refer this article as the law of heritage Cultural or, more briefly, as the law of heritage. It is a regulation aimed at developing the key elements of the heritage, and which designates as explicit objectives with respect to this acquis, the provide means of safeguarding and protection, recovery (solution to harmful trends or risk measures ), sustainability (co-financing that now allow the presence of assets and manifestations and mechanisms in the future) and disclosure (knowledge, recreation and people's access to assets and manifestations and their contexts).

This doesn't only mean that you must be able to appreciate it and enjoy it anytime, but it also means heritage, it should become an engine of economic and social development of the communities and should primarily affect positively in the quality of life of all Colombians who through projects and programs like however, it is necessary to point out that the law on Cultural Heritage focuses and develops a special regime of protection, safeguard, sustainability, in an integrated manner disclosure and stimulus to the real and personal property of that heritage which by its special values acquire, by using predefined procedures, the category of goods of Cultural interest, for which the acronym BIC can be used hereinafter.

Also focus on such a special regime in the demonstrations and activities of intangible heritage [sometimes called intangible, even though it's a discussion term] to be incorporated, given its special characteristics, and through a complex network of stages, to an instrument referred to representative list of Intangible Cultural heritage, for which the acronym LRPCI can be used thereafter. The special regime which is given to the goods of Cultural interest is called a special regime of protection, and special scheme for safeguard which is projected onto the demonstrations incorporated into the representative list of Cultural Heritage Immaterial, which involves the immediate adoption of a Special Plan for safeguarding - PES.

Llorenç Prats [16], defines the term heritage as those elements that represent social, cultural, and historical events which have generated an impact in a collective or social group considering not only the importance of their conservation but also the form is transmitted from generation to generation. Similarly heritage is considered is a social construction based on different ideologies, customs, experiences and stories [16,17]

## **2.3 Serious Games for Cultural Heritage**

From the concepts presented by several authors, is considered the existence of a relationship between the serious games and the cultural heritage, according to Mortara *et al.* [18], deduces that all serious games for the architectural heritage and natural must include puzzles and puzzles, proposed including the gamification and playability that facilitaria understanding of issues, taking into account that the puzzles are intended to collect, combine and use objects allowing interact with other people, these games are enriching because the player meets with an approach of "learning by doing", i.e., the player learns through the construction of their knowledge to meet a significant activity in this approach the Player does not passively receive information but that is actively constructing new knowledge through it in moving the plot of the game.

Moreover Rhim *et al.* [19], concludes that, without heritage, the identity of the community is at risk; Therefore, the preservation of the cultural heritage is important within the social development. One of the best ways to preserve the legacy is to increase awareness of its importance for some people learn about heritage could become tedious if traditional techniques are applied which involves the protection of heritage is difficult or requires much effort. This reflection suggests an easy and entertaining method of for young adults. Serious games are an interactive way of transmitting learning [20]. For example the keepers is a serious game designed to be played on mobile devices. In accordance with Smartphone property report from 2013, 156 percent of American adults own a Smartphone of some kind. As indicated in the report, many people have mobile devices and research shows that mobile technology offers learning in a natural environment [21].

Therefore, playing serious games on mobile devices has the opportunity of learning. With the incursion of these technologies is would have the possibility of improving the process of learning about the cultural heritage in search of increasing awareness of people improving the willingness to participate in the preservation of threatened heritage.

### 3 Methodology of the Patrimonial Serious Games Review

The methodology proposed for the development of the work is composed of five phases: (1) identification of search, (2) selection criteria and classification of primary education, (3) develop a protocol for review, and (4) analysis of the results (see Fig. 4) is presented below, the description of the various phases of the methodology.

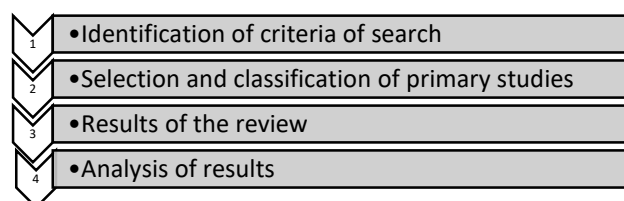


Fig. 4. Phases of the methodology processes. (Source: author's own).

#### 3.1 Development of the Methodology

In the process of investigation into the serious games focused on cultural heritage in Colombia search work was divided into five phases:

##### Phase 1: identification of criteria of search

A. search for scientific articles, PDF, books related to "Categorize + Serious Games + Heritage" search criteria.

This first stage refers to the identification of relevant games to the theme "cultural heritage" by conducting a search of scientific articles on serious games accessible through the World Wide Web., this stage consists of the following steps:

Search. to identify the set of lead articles: documents have been selected based on the key words "Categorize + Serious Games + Heritage". In the process of digital

databases search sheds large numbers of results, with those numbers take the first 20 and checked if these are consistent with the line of research that is developing. Below is a more specific search for the games in Colombia, thinking that some developments only count with the publication of the final product but there publications in revisstas and books.

**B.** search for Colombians serious games for cultural heritage "Categorize + Serious Games + Heritage".

Articles results: after the filter by words, reading abstract and introduction are deliberate and get 2 items correct and helpful for the construction of this article about Colombian serious games for cultural heritage; for this stage, it consists of the following steps:

- Identify the set of serious games: this is needed to have clarity of the key concepts for the search on web, trying to focus on the Colombian articles.
- After deliberating for names and location are obtained 3 articles that are in line with the theme of Colombian serious games for cultural heritage. We did a search pages directories of games development companies using as a basis a list of Colombian companies, software developers and was a filter according to the category of execution, this information can be found in table 2 Relationship of games found in Colombia, which presents important for the evaluation of the games features such as: availability, type of video game, durability and a brief description of it.
- For a correct selection took into account the different digital such as Google Scholar, DOAJ, IEEE, SciELO, OARE databases and other 30 bases free access on the internet, where the exclusion and inclusion criteria were selected the articles relating to the serious games for heritage project

**Phase 2: selection and classification of primary studies.** Once carried out the searches, are proceeded to select the material that contains important information and useful for the subject that is working, how long the process will use a reference management tool, which in addition to removing duplicates it provides access to any of them and generates a repository of references which can be consulted at any time. The time of execution of this phase was three months, used digital as SienceDirect, SciELO, ACM and IEEE among others databases.

Other collected information, was held also a classification of products was published in the application stores mobile and developments published on the internet for use without support of scientific literature, this genre also proximity with certain companies that develop such products, generating information value to the project.

**Phase 3: results of the review.** As the search results strengthens several examples of games and business developers in Colombia in the table 3 games found in Colombia, where the objective of the game is mentioned.

**Table 3.** Relationship of games found in Colombia.

Web site	Name	Availability	Types of video games	Information
	Museum Santuario de las Lajas	Yes	Educational arcade	It tells the story of the shrine of the Virgin of las Lajas and the apparition of the Virgin. The player must complete the proposed challenges during which is knowing the place and pick up the rewards.
12hitcombo	Santander	Yes	3D interactive virtual tour	Santander tourist is an interactive virtual tour in 3D, which presents 30 of sites representative of the Department of Santander, in Colombia, where the user explores and learns to close to them as you use the application.
Colombiagames.com		If for iOS 4.3 or later Compatible with iPhone, iPad iPod touch ad		The purpose of this application is to explore ecosystems and the creative culture of different indigenous ethnic groups of Colombia. Its aim is to introduce children to the mystic and natural world of the ancestors, using games with games.
Colombiagames.com		If for iOS 4.3		They narrate the indigenous stories where looking for users to know the ancient knowledge
Colombiagames.com		Yes for iPhone		Travel from different parts of the Colombian territory, looking for objects that have a deeper meaning in the myths and legends of the country.
Dinomotion.co	Heroes of the fatherland: The Pola	Yes [Android and PC]	Fighter	It refers to the historical fact of the independence of our country, in which re tells the story through the Pola, heroine of independence.
Dinomotion.co	Roots of Armero	Yes [Android and PC]	Virtual tour	It refers to one of the tragedies that has marked Colombia, in which in addition to a tribute to the victims recreates the architecture of the town erased by a natural disaster and thus their traditional stories.
Dinomotion.co	XTUR	Yes [Android and PC]	Virtual tour	X-Tur your Mystic offers a tour pro's Honda - Tolima. Known for its beautiful architecture.
Plantamer.com	Balam	Yes [Android and PC]	SG heritage	It presents different games and virtual tours that provide not just from the safeguarding of the colonial heritage but also to the preservation of the environment in Colombia
	TikTok			
	Verdetopia			

**Phase 4: analysis of results.** To conclude, at this stage, it is concluded that in Colombia several heritage serious games have been developed unfortunately some of the projects are not published and have a purely academic scope, which aren't available on platforms, as way there is no scientific publications and nor has its impact measurements, what is could achieve generating availability to be used in schools, universities, tourism, history, geography education where they take advantage of these resources. From the information contained in table 1, was conducted a review of companies with a response of the at where is allowed to obtain interviews with some

of them, highlighting the main features in the development of serious games whose objective is the dissemination of the heritage.

Initially the interviewees spoke about their experience of creating serious games for the Colombian cultural heritage, exponenting that his main goal is to tell the stories of Colombia in a different way, without losing its authenticity, from which is emphasized the motivation that produces focus digital products to cultural heritage taking to another level the dissemination of the same, usually showing the history and culture of small towns, or villages disappeared by weather or disasters natural and its impact not only in the history of the country but also in the wealth of their traditions. The companies agree on the importance of "make known our stories", not only those Reals that occurred at times historically known as the conquest and the colony, but also those that belong to the oral traditions transmitted from generation to generation.

With respect to the target audience, in general the idea is to be school-age children, although any user is able to use the game, focus on this population and therefore in educational institutions, seeking sponsorship in entities NGOs such as the ministries of culture, the municipal authorities and the departmental governorates which foster the dissemination of information relating to the heritage or in some cases by looking for resources, this last aspect generates disadvantages for developments.

As for the team, it is necessary to have professionals in different discipline, although the selection of members is closely related to the type of product that is expected to develop, in general educational equipment is required [historians, pedagogues, teachers of history and geography], programmers, graphic designers, animators, illustrators and experts in project management. In the majority of cases, firms hire part of the members of the team project, however, in the case of companies that already have a track record, the company's permanent staff is considered. In some cases, these people move to the geographic sites in which were the facts that you want to count on designed plays, in order to carry out research respects the facts, characters, and situations that surround it.

Within the relevant elements that highlight companies is that the construction of the story is considered of much importance, this consistency with the name and the design of the game, and the inclusion of the mechanics and dynamics of the game adapted to the story. These elements are fundamental at the time of the design, construction and development of the game.

## **4 Conclusions and Future Work**

Heritage serious games developed by Colombian research groups and companies appear to be interesting information of the architectural and cultural riches of our country, the oral traditions that are handed down from generation in generation and culture of the Colombian people, in a way attractive, interesting, striking and innovative, showing the talent that you have for this kind of developments aimed at safeguarding and dissemination the heritage and history of the country.

While there is a great potential in games seriously with respect to disclosure of the heritage and its care, exist in Colombia companies engaged in development this types of games however details of their products and measuring the impact that these they generate in the society and the public goal of their products is not found, despite this there is a great interest of the Government generated calls as MINTIC, creates Digital,

App, Digital living, who financed the development of this kind of applications, however requires that these applications are actually available to the greatest number of people, massive and not only in programs such as computers for education, where the percentage of people covered is very small, underused potential of these resource

Has begun a process of exploration of these types of games around the world, you will allow to establish as the development and dynamics of games, at what level this Colombia to other countries, than disclosure strategies are globally , and work can be done to increase the impact of these games in our country.

In addition, a guide of the characterization for the construction of serious games in the dissemination of the heritage and culture, projecting this work as basis for new ideas in creating serious games, is ahead at this time where would have a check list to the objective of the game is reached.

## References

1. Larez, B.: Emotional stimulation of the videogames: effects on learning. Magazine electronic theory of the education and culture in the information society, vol. 7. No. 2 (2006)
2. Bravo, C.: Marketing of guerrillas in the WEB 2.0 (2017)
3. Definición MX. <http://definicion.mx/juego/> (2017)
4. The traditional children's games, loss of values, University of Barcelona
5. Malajovich, A.: The game in the level initial in travels teaching in the education initial. Paidós (2000)
6. Serle, P.: Taught in key game. Linking games and content, No. Noveduc, Buenos Aires, Argentina (2008)
7. Romero, V., Gómez, M.: The playground and its methodology, Barcelona Altamar (2008)
8. LUDUS: What is serious Gaming (2017)
9. SGU: What are serious games (2011)
10. Aunt Transformation in Action, learning: the teaching, the learning electronic to the games serious. What method to use? (2017)
11. Game Marketing – Gamification.: <http://www.gamkt.com/2011/12/19/teoria-de-juegos-juegos-serios-y-gamificacion/> (2017)
12. Seriousgames: how to learn in to virtualworld. Université Paul Cezanne III-, Marseille, France (2010)
13. Genero-Bocco, M., Cruz-Lemus, J., Piattini, V.: Methods of research in the software engineering, RAMAL (2014)
14. CGR e-learning: ¿Qué es Edutainment? (2018)
15. Ministry of culture, Mincultura: Patrimonio cultural mueble (2015)
16. Prats, L.: Politics and society, Madrid: Universidad de Barcelona (1998)
17. Garcia, J.: On the Cultural heritage, inédito (1992)
18. Mortara, M., Catalano, C., Bellotti, F., Fiucci, G., Houry-Panchetti, M., Panagiotis, P.: Learning cultural heritage through serious games (2017)
19. Rhim, J., Yeom, H., Kim, M., Lee, S., Kim, J., Doh, Y.: The keepers: game design seriously to improve awareness of the preservation of the cultural heritage.

20. Deterding , S., Sicart, M., Nacke, L., O'Hara, K., Dixon, D.: Gamification: Using Game Design, ACM Digital Library (2011)
21. Schwabe, G., Goth, C.: Mobile learning with mobile game: design and motivational effects, Vol. 21 (2005)



# Proposal of an Interconnection Management Model and Availability of Internet Access for Things (MGID-IoT)

Chadwick Carreto, Jhovanny Pasaran, Alfonso Fernandez

National Polytechnic Institute, School of Computer Science, Mexico

ccarreto@ipn.mx, jepasarang@gmail.com,  
alfonso.fernandez.v@gmail.com

**Abstract.** This paper proposes a Model for the Management and Strengthening of Quality of Service (QoS) in the aspects of Interconnection and Availability of Access for networks of nodes and their application in the Internet of Things, it is a Model of Connectivity that will allow the services provided by a network of Mobile Nodes to be most of the time available to users when they require it. This is done through a layered model that implements between each layer a Quality of Service mechanism that allows in the first instance to measure the performance in that layer and secondly to establish an improvement or correction action according to the layer. The proposed Model aims to be endowed with a standard character to be implemented in different Services Architectures, which will allow optimizing the aspects of availability, connectivity, reconnaissance and monitoring of the state of the network, link between nodes and user connectivity.

**Keywords:** quality of service (QoS), internet of things (IoT), interconnection, availability, node networks.

## 1 Introduction

Mobile networks are evolving to different types of applications, providing new services to users and supporting new communications protocols, it is necessary to have Quality of Service (QoS) architectures that guarantee reliable data delivery, in time and form [1].

The subject of QoS research in mobile networks has been of great interest and models and architectures have been proposed that aim to provide a solution to this problem, however, there are still few model options that are compatible for the Internet of the Things (IoT). It is for this reason that the present work proposes a new layered QoS model designed to be compatible with any type of new generation mobile network and applicable to IoT schemes.

This QoS model is based on different metrics for the interconnection and network model according to each layer. It proposes to analyze and review the layers of: Applications, services and communications. Metrics will be done with existing and free tools in order not to generate a consumption of extra resources in the network. This model is implemented in a mobile node architecture in order to validate the model by analyzing the information and data generated.

The following sections will describe the state of the art of the networks and the aspects of QoS and from Section 4 will describe the proposed model. Finally, a series of conclusions and future work will be made.

## **2 New Generation Networks**

New generation networks have become a topic of great interest not only academically but also in the industry. These networks are characterized by high transfer rates to support various applications [2]. As time has passed the mobile networks have undergone changes in their architectures and also in the way they process information, these changes are called: generations of mobile communications.

One of the main characteristics of each generation is the way in which the mobile device accesses the channel or medium in the network, for example for a third generation network (3G) a code division medium access scheme is used (CDMA), while the fourth-generation (4G) network uses the Orthogonal Frequency Division Multiple Access (OFDMA) technique. For the fifth generation (5G) networks there is still no definite, so far the research done has simulated networks that could be considered as candidates for a 5G network. Another aspect that is important to consider is how QoS is performed in each of these generations [3].

## **3 QoS in Next Generation Networks and IoT**

Mobile communication systems base their QoS architectures on three main focuses [4]:

1. **Best Effort:** It is the model applied to any network that does not have policies explicitly defined as the Internet. It does not guarantee any treatment or specific resource to any information flow. Every package is treated equally; there is no preferential treatment.
2. **Integrated Services:** Implementation model of low-demand service that aims to guarantee resources available along a route for a specific application. Before starting the application session properly, the route is signaled to verify the availability of the necessary resources for an adequate development of the same. It allows to guarantee the operating conditions of critical applications.
3. **Differentiated Services:** It is a model of implementation of resources guaranteed in generic way and not by flows or sessions. It allows to guarantee different service conditions for different types of traffic, in a much more scalable and effective way, throughout the network.

The resource allocation is done jump-by-jump on each device in the network and not for a specific route. However, the implementation mechanism is relatively complex.

QoS services in IoT are considered to be end-to-end, meaning that they go from one end-to-end device, for example a mobile device connected to another through a network wireless [6].

Each end-to-end service has its own QoS which is provided to each user within the network, therefore the user is the one who can decide whether or not he is with the assigned QoS. To guarantee the QoS in a network, a bearer service with defined

characteristics and functionalities is adjusted in the network from the source of the service to the final destination [7].

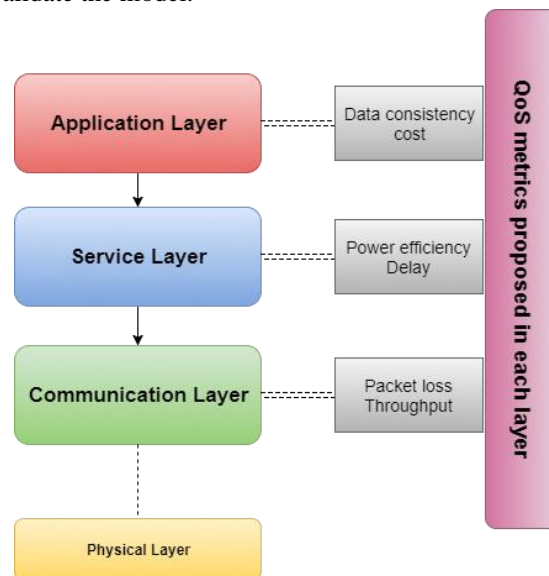
Each service carrier includes all aspects to enable the provision of contracted QoS. The QoS architecture in IoT is a layered architecture based on service bearers, where each bearer in a specific layer offers an individual service provided by the lower layers.

#### 4 QoS Model for IoT

One way to solve this problem is through a layered QoS model, the metrics will be made based on the parameters that each layer has. Fig. 1 shows the proposed model of this research work. The main characteristic of this model is that it will be developed in three layers:

As can be seen in the figure this model consists of three layers where QoS metrics will be made. In the application layer will measure which applications are most used. For the services layer, the types of services offered on the network will be obtained, ie whether they are real-time or not. Finally, in the communications layer, the metrics will be performed for: jitter, latency, bandwidth, packet loss, etc.

All metrics will be performed using existing tools in a network architecture of mobile nodes that will validate the model.



**Fig. 1** Layer QoS model proposal for new generation networks

Applications Layer. It refers to the metrics of the application levels. Some classification schemes for applications that require QoS levels [5] can be presented as follows:

- Elastic Applications: These are the ones that can adapt their operation and quality parameters of the services on the "best effort" scheme. In these applications we do not see a very close human interaction with communication. Examples are the

transfer of archives, or sending mail. Technologically, these applications do not require special parameters such as a specific bandwidth or a delays limit. Since in the case that they have a small bandwidth, it will only take more time to be sent, the transfer time will be longer or even exist mechanisms of data recovery and lost packets, correction of errors that are characteristic (but insufficient for other services) of the networks.

- **Non-Elastic Applications:** In contrast to previous applications these applications do require that they be guaranteed some QoS parameters. Real-time applications, for example, require bandwidth and a minimum level of delays for transmission stability. Although some of these applications show some ability to adapt to changes in quality parameters of services. (Without fail to ensure the same).
- **Interactive Applications:** Typically, a human interacts in this communication scheme, usually on an endpoint device that requires a fast response so that another user, device or host node can perform well. The parameters that are generally required to guarantee a high quality of services are strict limits for a low delay time, minimizing the error rate etc. Examples of these applications are: Voice over IP (VoIP), videoconferencing, online collaborative applications, online games, and so on.
- **Non-Interactive Applications:** Human direct collaboration is not required here to perform successful transmission. Nor are strict quality parameters required in the service. For example, you can minimize the low bandwidth with previous storage of the files or components. If audio or video streaming is required, it can be done with a slight delay (without storage) without this meaningful problem. Examples of these applications are: WEB browsing, file transfer, chats or preloading of audio or video content (streaming).

**Services Layer.** It refers to the metrics of the services offered by the network, see Table 1.

**Table 1.** Services can be classified according to [6].

SERVICES
Real Time
Signaling and Control
Data
Best Effort

- **Signaling and Control Services.-** This type of service is intended for traffic that supports network control, such as routing protocol traffic, and traffic signaling for voice and video , in particular this class must have a guaranteed bandwidth for its correct operation.
- **Data Service.-** This class of service will be designed for the highest priority data such as data bank transfers, FTP / SFTP services, etc.
- **Best Effort Service -** This kind of service will be devoted to data or video traffic that is for entertainment or game-oriented. Some applications for example are: mail servers, online games, music and video on the web, etc.

**Communications Layer.** It refers to the metrics in the communications channel as: latency, jitter, [6], etc.

- **Latency.-** Refers to the total time that elapses since a data packet is transmitted from a source node until it is received by the destination node, this parameter is measured in units of time. In real time applications such as video and voice it is necessary that there is a minimum level of delay in order to obtain a good quality of the application.
- **Jitter.-** In telecommunications jitter is called the variability of the execution time of the packages. This effect causes some packets to arrive too soon or too late to be delivered on time.
  - The jitter or delay variation can be caused by different factors such as:
  - Different packets may have different deadlines in queues on the same network device.
  - Different packets may have different processing times on the same network device with different delay times.
  - Different packets can travel through different network paths and delays would accumulate at different queue times and propagation delays.
  - From the perspective of users, jitter or lack of it, has repercussions on the consistency or consistency of the applications. However, while it is a low and constant level, applications can be adaptive.
- **Packet loss.-** It is the measurement of packets that have not been transmitted successfully on the network in relation to all packets sent on the network. Usually detected via ARQ (Automatic Repeat-reQuest) methods, there are four main causes of losing packets on the network.
  - Due to poor media quality either through physical or electromagnetic interference (often on wireless media).
  - Due to the congestion of links causing buffer overflow on the used network devices.
  - Faults in network devices.
  - Changes in the routing scheme or network protocols causing loss and damage to packets.

The loss of packages is reflected in the quality of presentation of the applications, such as: a good sound in an audio or sharpness if it is a video image.

## 5 QoS Model Test Architecture

The Quality of Service model was tested in the following non-two mobile architecture that aims to provide security monitoring in the halls, laboratories, offices, etc. of the National Polytechnic Institute. See Fig. 2.

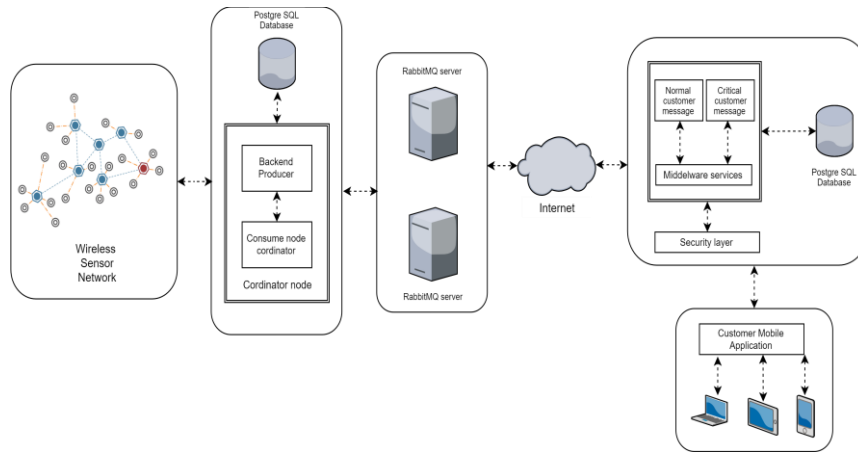


Fig. 2. Mobile node architecture proposed.

Architecture is divided into five important parts: 1) Xbee Network; 2) The coordinating node; 3) RabbitMQ Servers; 4) The middleware, and 5) The Web application (Client).

**Xbee Network.** The xbee nodes will be located at the doors of each school and through a proximity sensor they will be monitoring if any doors open at an inappropriate time. As shown in Fig. 3 there must be a coordinating node to which the other nodes are connected, and this node will be the gateway that will distribute network traffic to the next part of the architecture.

**Node Coordinator.** The coordinating node consists of a small server that will be in charge of receiving all the information that is being censored by the whole network of Nodes, so there will be a module called "Consumer Node Coordinator" that will have two important functions the first to perform the register of all the information of each one of the nodes, this will be done by means of a PostgreSQL database where the detailed information of the node will be stored (serial number, variable to be census, Possibly the frequency in which it is located, location, active ) so that it will serve to receive each of the configurations by the user (Customer), the second function will be to receive all that data and then send them through the "Ba-ckend / Producer / Consumer". In the case of "Backend / Producer / Consumer" will be responsible for establishing two types of communications:

- **Producers:** It will establish two communication channels, one to send the data that are sent by the nodes and another channel to be able to send messages of high priority which can be sent given several situations (one node was decompressed, the base of data stops working).
- **Consumers:** In this case, a single communication channel will be established, which will be responsible for receiving the configuration data of the nodes, that is to say that some of its operation can be altered, such as to disable it, give it even priority or simply to return location).

**RabbitMQ Servers.** This part of the system consists essentially of the servers that will be raised by RabbitMQ which are mainly the "Tracker" that will be in charge of redirecting each one of the messages to its different destinations, in this case two or more are proposed (depends on the number of schools required) that will have two main functions:

- **Server 1 (Important):** This server is very important because it will be responsible for redirecting those messages that are very important, when there are already high-impact situations (a node was shut down / decomposed, the data base left to operate).
- **Server 2 (Data):** This server will mainly be used to manage all the information that will be registered by the nodes.

**Middleware.** This middleware is a fundamental part because it will take care of receiving / sending all the messages that are necessary for the system, this middleware is composed by 5 parts:

- **Consumers:** There are 3 connections that will remain connected to the servers in order to be able to receive all the data that is sent from "Backend", specifically two are required so that the work can be distributed in both "Consumers" and thus any of the two stacks are saturated, although the connection is more focused to the channel where those alerts or problems will be sent by the coordinating node.
- **Producers:** This will be in charge of sending all necessary information on the configuration of the nodes, so it is necessary only a single Producers since it is a task that will not be used continuously.
- **Sails Services:** This is an MVC framework made for Node.js that will help us primarily to provide all necessary API Rest services to be able to interact with the client, but in turn to empty it into a NoSQL database, as in the case of the CassandraDB database, so you will be interacting directly with the 3 Consumers and the Producer in order to send / receive information.
- **Security (Authentication):** It is a layer that will be responsible for establishing a security layer between the SailsJS services and the Client, this will be done through an integration that has the same SailsJS Framework in this case the pro-tocol that will be used , OAuth 2.0 that consists of the handling of tokens that are valid for a certain time and that are used to be able to accede to the SailsJS API, reason why only valid users will be able to request to the server a valid token and with only certain privileges.
- **CassandraDB database:** This database will store all information related to the system, ie profiles, users, tokens, etc., including the same information of the nodes, this CassandraDB case is considered for solve problems of scalable systems, so it will be useful when handling a large amount of information.

**The Web application (Client).** This application will be made using Angular 2, considering two important aspects, the first is the robustness that will be necessary for the system so Angular 2 along with TypeScript will be a great help to be part and the second is that it allows us to have greater benefits for the case of mobile devices, even give an experience to the user that it is a mobile application when it is a web application.

Here are the possible profiles that the client application will have:

1. User Manager.
2. Node Manager.
3. Monitoring User.

Where the same system will contain different modules:

1. Module for User Administration.
2. Node Management Module.
3. Notifications module (to warn users of major incidents).
4. Module for generating reports.
5. Module for monitoring the different nodes and variables.

## 6 Conclusions

The main reason for proposing and implementing this model is that it can become a flexible and robust medium or method that, at the moment of being implemented within the context of a mobile network or an IoT system, allows users to have an "Acceso" to the corresponding information, services and / or applications wherever and whenever they need it, in a more efficient and reliable way. The Model is not intended to work with a specific access technology, however for our case study we will choose to work with a single IoT Access Technology. At the moment and with what it has mounted has proved to be a viable option with multiple application options.

## References

1. Viswanathan, H., Weldon, M.: The Past, Present, and Future of Mobile Communications. *Proc. Bell Labs Technical Journal* 19, 8–21 (2014)
2. Wunder, G., Boche, H., Strohmer, T., Jung, P.: Sparse Signal Processing Concepts for Efficient 5G System Design. *IEEE Access* 3, 195–208 (2015)
3. Andrews J. *et al.*: What Will 5G Be? *IEEE Journal on Selected Areas, Communications* 32(6), 1065–1082 (2014)
4. Agiwal, M., Roy, A., Saxena, N.: Next Generation 5G Wireless Networks: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials* 18(3), 1617–1655 (2016)
5. GSMA Intelligence: Understanding 5G: Perspectives on future technological advancements in mobile. White paper (2014)
6. Solomitskii, M.: About Quality of Service concept's usage at delivery of messages streams circulating in convergent telecommunication network. In: *Microwave and Telecommunication Technology (CriMiCo), 23rd International Crimean Conference, Sevastopol*, pp. 430–431 (2013)
7. Rajveer, H., Hemant, S.: Comparative Study of Different Mobile Generation Technologies and Various Challenges in Integrating 4G and 5G. *IEEE Communications Magazine* 42, 165–187 (2016)

# Architecture of Mobile Services Applied to the Internet of Things

Chadwick Carreto<sup>1</sup>, Francisco Cerda<sup>1</sup>, Felipe Menchaca<sup>2</sup>

<sup>1</sup> National Polytechnic Institute, School of Computer Science, Mexico

<sup>2</sup> National Polytechnic Institute, School of Mechanical and Electrical Engineering, Mexico  
ccarreto@ipn.mx, fco.cerda.mtz@gmail.com, fmenchac@gmail.com

**Abstract.** In this paper, we propose a service architecture that integrates an information exchange architecture with the objective of standardizing the communication between different networks of nodes that implement different protocols such as IEEE 802.15.4, Zigbee, DigiMesh, Thread. This is achieved by means of a correct management of the proposed frame for its interconnection, as well as attending different aspects of the communication and establishment of links, this for networks of nodes that allow the handling of information for applications and services applied to the Internet of Things (IoT). This architecture proposes to implement within a system of unification of various medical and academic services.

**Keywords:** internet of things (IoT), interconnection, exchange, standardization, architecture.

## 1 Introduction

With the arrival of IoT (Internet of Things) comes the technology of sensor networks which brings us a new problem, one where information and communication systems surround our personal and professional areas which brings with it the generation of enormous amounts of Information and the necessity to have a place to store the information, as well as one to present it, to exploit it and to treat it to an easily interpretable form.

In order for the Internet of things vision to work correctly, the paradigm of computing must go beyond the traditional approaches of mobile computing, in this paradigm we recognize that user is not the only user of a systems, and user generates more information that he believe (pervasive systems).

The IoT requires three primary factors which are desirable in most cases. A shared understanding of the situation of users and their devices. Perceptive software architectures and communication networks, allowing not only to transmit the usual information but also to include contextual information where necessary in order to provide better services.

The existence of analytical tools in order to achieve an autonomous and intelligent behavior of the network, as well as the services provided [1]. In addition to this, there is a need to employ some type of architecture and define a set of protocols and formats of exchange in order to achieve or address the points mentioned above. At present, there

is a set of architectures that allow the development of state-of-the-art technologies, some architectures are used in a particular way for specific systems [2].

There is also the possibility of using hybrid architecture to enhance the best aspects of each one. But this brings with it the problem of standardizing these structures to guarantee intercommunication and interconnection between them.

In the last decade, advances in Information and Communication Technologies (ICTs) have allowed the optimization of services that are used in fields such as industry, education, medicine, environmental protection, etc. Such is the case that in recent years it has been decided to make use of ICTs to automate the use of clinical records and thereby reduce the waiting time for patient care and improve the control of information. This is why it is necessary to use an architecture that allows the integration and optimization of various medical services so that they can be used by the right-holders, patients, doctors, researchers, administrators, managers, etc. which are part of the health sector [3].

The following sections describe the existing Services architectures and the proposal to be implemented to test their characteristics.

## **2 Communication Protocols**

The architecture is defined as the structure that defines which components must possess a system as well as the relationships between each one of them. The design of a software architecture represents one of the most critical decisions for the correct functioning of a system, since this architecture will define many of the characteristics that a system can achieve [2].

**IEEE 802.4.15 Protocol.** IEEE Std 802.15.4-2003 defined the protocol and the compatible interconnection for data communication devices that use low-power, low-power, low-complexity short-range radio-frequency (RF) transmissions in a wireless personal area network (WPAN). This revision expands the applicability of the IEEE Std 802.15.4 market, eliminates ambiguities in the standard, and improves IEEE Std implementations 802.15.4-2003

IEEE 802.15.4 was designed to operate in unlicensed radio frequency bands. (Normally, RF output envelope rules and possibly the duty cycle of a device operating in these bands are still applied). The unlicensed RF, bands are not the same in all territories of the world, but IEEE 802.15.4 employs three possible bands, at least one of which should be available in a given territory. The three bands focus on the following frequencies: 868, 915 and 2400 MHz. [4]

**ZigBee.** It is a wireless communications standard designed by the ZigBee Alliance. It is a standard set of solutions that can be implemented by any manufacturer. ZigBee is based on the IEEE 802.15.4 standard for wireless personal area networks (wireless personal area Network, WPAN) and targets applications that require secure communications with low data rate and maximize the life of your batteries [5].

The ZigBee protocol defines three types of nodes: Coordinators, Routers and Final Device, with a requirement of a Network Coordinator. Although all nodes can send and receive data, there are differences in the specific functions performed by each of the different types.

Coordinating nodes are the most capable of the three types of nodes, there is a coordinator per network and it establishes the origin of the network is able to store and/or manage network information, including security information.

Nodes Coordinators.- act as intermediaries between nodes, and transfer data from other devices.

Final Device.- Those devices that have sufficient capacity to speak with higher hierarchical nodes (roots) and cannot communicate with other devices, so as to reduce cost are seen with diminished functionality [6].different rates of data. [4]

**Digi Mesh.** Digi Mesh is a protocol that only uses a node type, which generates a homogeneous network, this means that all nodes can route information and are interchangeable since they do not have a relation to their root nodes. DigiMesh is a proprietary networking topology for use in wireless endpoint connectivity solutions. It supports advanced networking features, including sleeping routers and dense mesh networks. DigiMesh supports multiple network topologies such as point-to-point, point-to-multipoint and mesh networks. With support for sleeping routers, DigiMesh is ideal for power-sensitive applications that rely on batteries or energy-harvesting technology [6].

**Thread.** It is an open standard for reliable, cost-effective and low power D2D (device to device) communication. It is specifically designed for connected home applications where IP-based networking is desired and a variety of application layers can be used in the stack. Here are the general characteristics of the stack of wires and the network:

- Simple network setup, commissioning, and operation: Simple protocols for forming, joining, and maintaining wire networks allow systems to autoconfigure and correct routing problems as they occur.
- Secure: Devices do not join the subnetting network unless they are authorized and all communications are encrypted and secure.
- Small and large networks: Home networks vary from several devices to hundreds of devices that communicate transparently. The network layer is designed to optimize network operation based on expected usage.
- Range: Typical devices along with mesh networks provide enough range to cover a normal home. Extended spectrum technology is used in the physical layer to provide good immunity to interference.
- There is no single point of failure: the battery is designed to provide safe and reliable operations even with the failure or loss of individual devices.
- Low power consumption: Host devices can normally operate for several years on AA-size batteries using suitable duty cycles [7].

### **3 Middleware**

Middleware is software that sits between an operating system and the applications running on it. Basically, it works as a hidden translation layer to allow communication and data management in distributed applications can be referred to as “plumbing”, as it connects two applications so that data and databases can be easily passed through one “canalization”. The use of middleware allows users to make requests such as submitting

forms in a web browser or allowing a web server to return dynamic web pages based on a user's profile [8].

A Middleware can be seen as a set of reusable, expandable services and functions that are commonly used by many applications to function well within an interconnected environment [9].

Some common examples of middleware are database middleware, application server middleware, message oriented middleware, web middleware, and transaction processing monitors. Each program typically provides messaging services so that different applications can communicate using messaging frameworks such as Simple Object Access Protocol (SOAP), Web Services, Representational State Transfer (REST), and JavaScript Object Notation (JSON). While all middleware performs communication functions, the type a company chooses depends on the service being used and the type of information to be communicated. This can be security authentication, transaction management, message queues, application servers, web servers, and directories. The middleware can also be used for distributed processing with actions that occur in real time [8].

## **4 Services Architectures**

The Architecture is defined as the structure that defines which components must possess a system as well as the relationships between each one of them. The design of a software architecture represents one of the most critical decisions for the correct functioning of a system, since this architecture will define many of the characteristics that a system can achieve [2].

**Service Oriented Architecture.** Mental elements for the development of solutions, a service will be described as a computational element agnostic to the platform that supports a fast and low cost, that make up distributed applications. Services perform functions, ranging from simple requests to complex business processes. Services allow organizations to expose their core competencies programmatically over the Internet (or intranet) using standard languages and protocols (based on XML, JSON or some other exchange format), and implemented through a self-describing interface based on open standards [10].

**Architecture PipeLine.** A pipeline connects components (filters) through connectors (pipes), so that the data is processed and executed as a flow. The data are transported through the pipes between the filters, gradually transforming the inputs into outputs [11].

**Architecture Peer to Peer.** We can consider a P2P system as a multilayered architecture, namely: The network level represents the lowest level of the architecture, offering the basic capabilities of communication between computers, either through IP networks or through Ad-Hoc networks. The level of administration of the P2P layer, responsible for routing messages and performing overlay maintenance tasks. The level of features, which supports network functionalities such as security, fault tolerance or resource management. The level of services, which is in charge of providing functionalities at the application level [1].

**Event Oriented Architecture.** Architectures with the following characteristics:

- Broadcast communications. Participating systems broadcast events to any group. More than one party can hear the event and process it.
- Opportunity. Systems publish events as they occur instead of storing them locally and await the processing cycle, such as a Discontinuous Cycle.
- Fine grain events. Applications tend to post individual events rather than a single aggregate.
- Ontology. The global system defines a nomenclature for classifying events, typically in some form of hierarchy. Receptor systems can often express interest in events or categories of events.
- Complex Event Processing: The system understands and monitors relationships between events, for example aggregation of events (a pattern of events involves a higher level) or causality (one event is caused by another) [12].

**Three Level Architecture.** Also known as Three Tier, or Three Scheme Approach. The purpose of the architecture of the three schemes is that:

- All users should be able to access the same data
- A user view must be immune to changes made in other views
- Users should not know the physical details of database storage
- DBA must be able to change the storage structures of the database without affecting the view of the users
- The internal structure of the database should not be affected by changes [12].

## **5 Mobile Services Architecture for IoT**

After analyzing the current solutions and after reviewing and analyzing the existing communication protocols of the nodes networks. It seeks to determine in detail the desirable characteristics of the different modules of the proposed architecture, in order to define the necessary elements and the necessary processes within the different modules of the architecture, in order to later implement those modules, once these modules have been developed separately.

**Definition of architecture modules.** The modules proposed so far will be shown and explained in general below. The proposed architecture will have at least 3 main aspects which are the interconnection module which will be a middleware that will allow the interchange of the communication between the different sensor networks and the service connection architecture.

To achieve this, it is necessary to establish a common frame format which is operable by the hardware and software middleware in order to be able to connect with the service architecture.

Finally, the service connection architecture will be responsible for consuming and exploiting the information generated by the different sensor networks as well as providing services (functionalities) according to the user profile and its requirements. See Fig. 1.



Fig. 1 System architecture.

## 6 Design and Development of Architecture

The SSA modules are described below:

- **Administration and Configuration Module.** This module is aimed at verifying the work of each of the other modules, as well as updating the databases in terms of configuration of devices and IoT.
- **Services Module.** This module is located both in the client and in the server, it has the functionality of providing and requesting the services and the handling of the information stored in the databases interacting with the user. These services can only be used by authorized users and can view the information they require according to the type of user.
- **Connectivity Module.** This module is one of the most important since it implements what is the connection and disconnection in IoTs between the server and the client, this module presents greater dependence in relation to the authentication module already that in order to establish the connection, it requires authentication beforehand.
- **Authentication Module.** This module is responsible for the validation of users who wish to establish connection, obtains the identifier and the address of the device whether the device assigns one automatically or is provided one in the context of the IoT.
- In this module in the client, the device is responsible for automatically finding a node of the available IoT network or that it is allowed to enter in order to be subsequently provided with the connection, authentication and can be provide the services.
- The authentication module is based on Oauth 2 that implements an access control based on tokens which are generated a refreshed by requests to server.

- **Device Identification Module.** As the architecture is independent of the devices used it is necessary to verify the capabilities of the device to be accessed and for this the server takes the device identifier, recognizes it and once the type of device was obtained determines how to provide the services in its respective module.

The architecture was tested by setting up a test case at the Higher School of Computer Science of the IPN, divided into a network IoTs a school control, a library domain and an Internet communication service, on different elements and with a central server of backup although it is not necessary for the architecture:

A system was developed that allows the members of the community of the National Polytechnic Institute to have an agile and comfortable access to the medical services offered by the institution, as well as a means of identification that establishes them as members of the community (students, teachers or employees). Likewise, the means in addition to identifying them will contain the vital personal and medical information of the applicant in order that those in charge of providing a service within the institute can offer it in a simple manner and also have a history of the activities of the applicant.

Each of the attention units within the institute where the services for the polytechnic community would be offered will be able to attend to the applicant requesting their card in order to obtain their information and identify him as a member of the IPN, this operation will be carried out through an NFC reader that is you will be connected to a desktop or wireless application where the information necessary to access the service will be displayed.

It will also have a server where the history of all the services that a person has requested and the history necessary for specific cases will be stored, being for example the history of medical consultations for the case of any medical area or the history of loans and debts for the library service.

The system will also have a web portal where users of the community will be able to access with a username and password in which they will have the opportunity to visualize the history of their requests and their general information.

## **7 Conclusions and Future Work**

The main contribution of the proposed architecture is to allow to act with more mobility saving time and effort in the access to the information. The system itself is another way to help people carry out their functions in any area they develop, not only because it facilitates the job but also because it opens a door to innovation and a contribution to society.

The IoT with mobile nodes are becoming more common in both institutions and companies, it has now become a goal to achieve a ubiquitous computation that ensures total interactivity at all times and everywhere, so this project is focused on getting closer to these objectives, aiming to develop an environment capable of providing the services required for specific users in the most transparent way possible.

The system could also be a good basis for developing a ubiquitous and IoT computing environment wherever it is needed, establishing a protocol for the implementation of services for mobile equipment of various types.

The application of this architecture can be in many areas of knowledge, including:

- **Medicine.** In this area, it can be implemented in a hospital where when a patient arrives at reception with his cell phone can make an appointment, or at the same time can obtain information from an inpatient.
- **Education.** One form of application is that they can obtain mini applications on their mobile device such as information services, searches, virtual laboratories, simulation systems, etc.

## References

1. Ashton, K.: That ‘Internet of Things’ thing. RFID Journal (2009)
2. Del Rosso, C.: Continuous evolution through software architecture evaluation: a case study. Journal of Software Maintenance and Evolution: Research and Practice 18(5), 351–383 (2006)
3. Diario Oficial de la Federación: Reglamento Interior de la Secretaria de Salud. pp. 8–40 (2016)
4. IEEE 802.15.4 Stack User Guide (2017)
5. Application Standards: <http://www.zigbee.org/zigbee-for-developers/applicationstandards/> (2017)
6. Wireless Mesh Networking ZigBee® vs. DigiMesh. White Paper. Digi International. [https://www.digi.com/pdf/wp\\_zigbeevsdigimesh.pdf](https://www.digi.com/pdf/wp_zigbeevsdigimesh.pdf) (2017)
7. Thread Stack Fundamentals. 1st ed. Thread Group, pp. 3–19 (2015)
8. Microsoft: What is Middleware - Definition and Examples: Microsoft Azure (2017)
9. Aiken, B., Strassner, J., Carpenter, B. *et al.*: rfc2768. Cisco Systems, IBM, Argonne National Laboratory *et al.*, *Ietf.org*. (2017)
10. Sundmaeker, H., Guillemin, P., Friess, P., Woelfflé, S.: Vision and challenges for realizing the Internet of Things. Cluster of European Research Projects on the Internet of Things—CERP IoT (2010)
11. Buckley, J.: The Internet of Things: From RFID to the Next-Generation Pervasive Networked System., Auerbach Publications, New York (2006)
12. G. Hohpe: Programming Without a Call Stack – Event-driven Architectures. 1st ed., pp. 2–3 (2006)

# Embedded System for the Regular Blood Pressure Monitoring

V. H. Garcia<sup>1</sup>, N. Vega<sup>1</sup>, R. Hernandez<sup>2</sup>

<sup>1</sup> Instituto Politécnico Nacional, Escuela Superior de Cómputo  
Mexico City, Mexico

<sup>2</sup> Instituto Politécnico Nacional, UPIITA  
Mexico City, Mexico

vgarciaortega@yahoo.com.mx, nvegag@ipn.mx, rhtovar@ipn.mx

**Abstract.** This paper presents the development of an embedded system for monitoring a patient's blood pressure, automating the inflation and deflation stages of the bracelet using an FPGA. The embedded system is based on the oscillometric method to get the patient's systolic and diastolic blood pressure. This method monitors the variations or oscillations of the pressure signal in the bracelet that is placed around the arm, to determine through the autocorrelation function the values of the above mentioned pressures. The system uses a digital pressure sensor for the data acquisition stage, an air pump and a mini solenoid valve. The system is implemented in the Xilinx's Microblaze soft-core processor configured in a Spartan 6 family FPGA. The Microblaze processor uses the processor local bus to communicate the peripherals. A IIC core is used to interface with the digital sensor and a GPIO core to control the air pump and a TIMER core to control the mini solenoid valve with a PWM signal. Finally, a UART core is used to send data to a personal computer where an application server is implemented. This server allows remote monitoring of pressure sensor information by medical staff.

**Keywords:** FPGA, embedded system; health care.

## 1 Introduction

The vital signs are clinical parameters that reflect the physiological status of the human organism, and essentially, they provide data which will set the standards for evaluating the homeostatic state of the patient, showing his actual health condition as well as the changes or evolution, either positively or negatively. Vital signs include: temperature, breathing frequency, heart rate and blood pressure, among others.

Nowadays, vital and biomedical signs cover a wide spectrum of applications in different contexts, so that it has been a recurring research subject during the last decade [1-5]. Currently systems designed for vital sign monitoring which use different technologies exists.

Some researchers [2,3,5] have proposed to use microcontroller based technology to monitor several vital signs like heart rate, breathing, among others. In [4] an ARM

processor based embedded system is used with a Linux operating system so as to monitor the electrocardiographic signal (ECG) of a patient.

The proposed system is an embedded system which uses a field programmable gate array (FPGA) to monitor blood pressure (BP). Blood pressure is the force the blood exerts against the arterial walls. Every time the heart beats, it pumps blood towards the arteries, producing a pressure that is measured in mmHg. This pressure is made up of systolic blood pressure (SBP), the first figure, which is the maximum value the circulatory system registers when the heart contracts to pump blood to the arteries and deliver it to the whole organism; and diastolic blood pressure (DBP), the second figure, which represents the minimum value the artery register when the heart relaxes to be filled with blood again.

Hypertension or High Blood Pressure (HBP) is a sustained elevation of SBP, DBP or both which affects a considerable part of adult population, particularly the elders. The HBP is defined by the continuous presence of equal or greater than 140 mm Hg SBP figures, equal or greater than 90 mm Hg DBP figures or both [6].

In care practice, there are several different measurement methods that can be classified as: capillary color change, through pulse, auscultatory, oscillometric, doppler, intra-arterial or forthright. Among these, the most used method is the auscultatory one given that it is a non-invasive external method, the most accurate one and the most studied and researched. Notwithstanding, during recent years electronic devices for BP measurement are being introduced in everyday clinic practice. Nowadays, these electronic devices are preferred due to their easiness, comfort of use and to avoid the observer bias (it is the observer himself one of the main sources of inaccuracy in BP measurement) [7].

The proposed embedded system uses the oscillometric method as a base to obtain the SBP and DBP of the patient. The difference with the previous one is that this method monitors the variations or oscillations of the pressure signal on the cuff that goes around the arm to determine, through the obtained signal analysis, the values of the aforementioned pressures. In other words, while the cuff is being deflated from a value above the SBP, the artery walls begin to vibrate or oscillate as the blood flows through the partially occluded artery and these vibrations are captured by the sensor that monitors the pressure on the cuff. As the pressure keeps dropping, the oscillations increase until a maximum amplitude and then they decrease until the cuff is totally deflated and the blood flow returns to normal.

The pressure on the cuff during the maximum oscillation point normally corresponds to the mean arterial pressure (MAP). The point above the MAP, in which the oscillations begin to quickly increase in amplitude corresponds to the SBP. The point in which the oscillations variation decreases most abruptly corresponds with the DBP.

## 2 Proposed System Architecture

The embedded system is constituted by three modules, which are:

- Data acquisition and control.
- Embedded system
- Data display software.

The system architecture is shown in Fig. 1. In the system, the patient's arterial pressure signal acquisition is done through a digital sensor afterwards, the signal is sent to a soft-core processor synthesized on a FPGA where is processed and sent to a data display software on a personal computer.

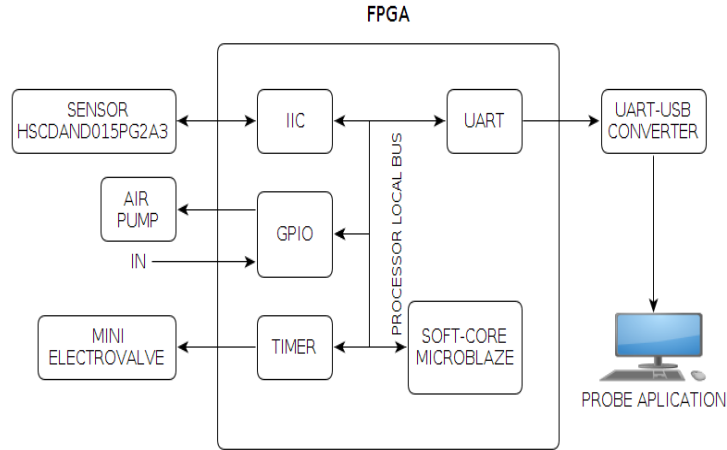


Fig. 1. System Architecture.

## 2.1 Data Acquisition and Control

The data acquisition module oversees obtaining the blood pressure signal of a patient in a digital for so it can be processed by the FPGA. This module is constituted by the following elements:

1.– Pressure digital sensor model HSCDAND015PG2A3 form Honeywell manufacturer [8]. It measures manometric pressure, also known as Gauge, is completely calibrated and compensated, handles a measuring range from 0 to 15 psi (equivalent to 0 to 775.72mmHg). Particularly, this prototype can detect blood pressure in a range from 0 to 400 mm Hg (equivalent to 0 to 7.7345 psi). It has a 14 bits analog-digital converter integrated, requires an input voltage of 3.3V and sends data through the Inter Integrated Circuit Bus (IIC) in the embedded system architecture. Furthermore, it has a total band error (TBE) of 1%, namely  $\pm 7.7\text{mmHg}$ . The sensor has the transfer function shown in equation 1. With this function, the pressure applied to the sensor can be determined based on the output value of the sensor:

$$Output = \left( \frac{Output_{max} - Output_{min}}{P_{max} - P_{min}} \right) * (P_{applied} - P_{min}) + Output_{min} \quad (1)$$

where:

$Output_{max}$  : Maximum output value (90 % de  $2^{14}$ ).

$Output_{min}$  : Minimum output value (10 % de  $2^{14}$ ).

$P_{\max}$  : Maximum supported pressure (15 psi).

$P_{\min}$  : Minimum supported pressure (0 psi).

$P_{\text{applied}}$  : Pressure applied to the sensor.

*Output* : Output value of the sensor.

2.– Standard size nylon cuff.

3.– Air pump that requires an input voltage of 6V and 460 mA of current. It handles a maximum pressure of 400 mm Hg, for the automatized inflation of the cuff stage. The relevance of the inflation stage lies in that is in this stage where the SBP of the patient can be immediately determined. This happens when the brachial pulse disappears, namely the artery is completely occluded by the cuff. The pressure value registered in the cuff by the sensor corresponds to the SBP. Nevertheless, due to time constraints, no algorithm to identify this event was made in the present work. For the time being the cuff is always inflated until 160 mm Hg followed by the deflation stage.

To handle the air, pump a General Purpose Input Output (GPIO) peripheral is used in the embedded system architecture. For this reason, a power stage was designed using a 4N30 optocoupler to avoid damaging the FGPA. The power stage is shown in Fig. 2.

Mini air electrovalve that requires an input voltage of 6V and a current of 250 Ma, it allows for a maximum escape speed of 88.3 mm Hg/s for the automatic cuff deflation stage. Given the case of the electrovalve the use of a pulse width modulated (PWM) signal is proposed to control the output air flow, ensuring the output don't go over the 3mm Hg/s ratio which is indicated by norm. To manipulate the mini electro valve a timer is used on the embedded system architecture. Taking that into account, is necessary to remove the optocoupler and use solely the transistor to obtain the current needed to activate de valve. This is because the optocoupler doesn't support an input current with a frequency higher than 1 KHz. From this frequency value, the optocoupler always provides a logical 1 as output, completely activating the mini electrovalve resulting on an immediate deflation of the cuff at a higher ratio than 3 mm Hg/s. The power stage is shown in Fig. 3.

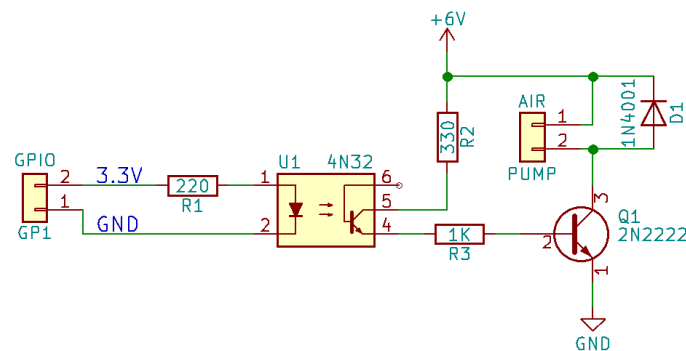


Fig. 2. Power stage for air pump.

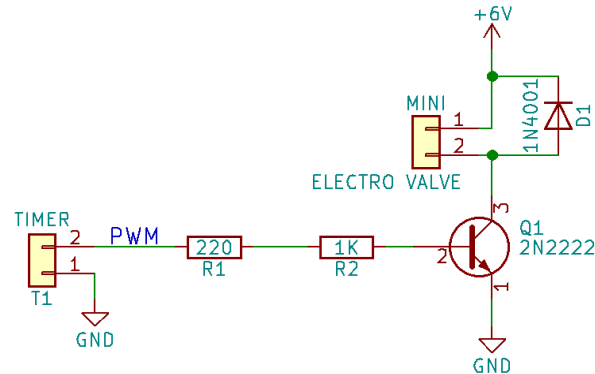


Fig. 3. Power stage for mini electro valve.

## 2.2 Embedded System

The embedded system has a Spartan 6 family, XC6SLX16 model Xilinx FPGA [9] as the processing unit with the Nexys 3 [10] development board from Digilent Inc. It has a 576Kb Block RAM memory, enough to embed the application on the Microblaze soft-core processor, responsible of all the calculations done in the system. The development board Nexys 3 offers its development environment (software) to configure the embedded system and imprint it on the board. The environment is the Xilinx Design Tools software that includes the Environment Development Kit (EDK) which on turn, is composed by Xilinx Platform Studio (XPS) to implement the necessary hardware configuration for the embedded system and by Xilinx Software Development Kit (SDK), which is used to link the hardware system created with XPS with the software application (usually in C/C++ language), thus creating a tailormade system.

The embedded system is formed by the following elements:

1.– Embedded system architecture: One of the prime advantages of working with a hardware reconfigurable FPGA board is that, as the name indicates, it allows the user to choose the hardware modules that he wishes to implement on his embedded system. The embedded system is constituted by the following modules:

- a. GPIO core Driver to control the activation and deactivation of the air pump.
- b. Timer/Counter core Driver in charge of producing the PWM signal to control the activation and deactivation of the mini electrovalve.
- c. IIC Driver to communicate the blood pressure sensor with the board. The sensor has an output of up to 4 bytes (data) depending on the options it has been configured with, according to the application needs. Under any circumstances, the first two bytes correspond to pressure values, while the third and fourth bytes correspond to a temperature value (that is optional to use). The sensor has an address equal to 0X28 and is configured at a standard speed of 100 KHz. The reading format is shown in Fig. 4.

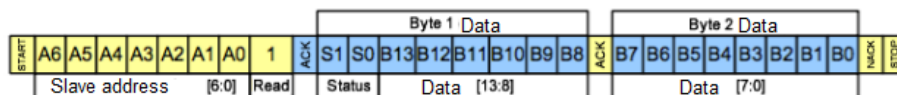


Fig. 4. Sensor reading format.

- a. UART Driver to send the data processed by the microblaze sof.core for its storage and subsequent consultation.
- b. PLB bus Driver that communicates the GPIO, Timer/Counter, IIC and UART modules with the microblaze softcore.
- c. Driver for the data and instruction memories DLMB and ILMB of the microblaze softcore.

### 3 Processing Algorithms

The main problem of the oscillometric method consists in the detection and amplification of the oscillatory pulses that are generated in the cuff during the deflation stage, starting from the data collected from the sensor, as well as the subsequent analysis of these pulses to calculate SBP and DBP.

There are algorithms sensitive to pulse pressure difference and to artery stiffness, however, most of them are well kept as commercial secrets, so that the exact SBP and DBP attainment using the surrounding curve of the oscillations that occur on the cuff during the deflation stage continues being an open problem in biomedic engineering [11]. Given that, we developed our own algorithm on an experimental way.

In this case the sensor recollects data during the cuff's inflation and deflation stages and sends it to the FPGA board at a transfer speed of 100 KHz. The cuff is inflated until it reaches a 160 mmHg pressure which indicates that, at the present time, the cuff can only measure the BP of patients that have degree 1 or lower HTA, this is due to the lack of an algorithm to determine when the artery becomes totally occluded during the inflation stage, the pressure when this event is detected corresponds to the SBP.

Using the collected data, the oscillations that occur on the cuff during the deflation stage are detected. In this case, this is done by the autocorrelation function calculation every 64 sensor reading, according to equation 2:

$$r_{xx}[n] = \frac{1}{N} \sum_{m=0}^{N-1-n} x[m]x[m+n] \quad (2)$$

$$n = 0, 1, \dots, N-1$$

The algorithm shown on Fig. 5 and Fig. 6 describes the pseudocode used in this system to determine the SBP and DBP of a patient. As input, it takes at most 3000 readings from the pressure digital sensor. The algorithm begins with the global variables declaration, an initialization stage of: FPGA buttons, GPIO bus and PWN pulse.

The algorithm on Fig. 5 starts with an infinite loop where an auxiliary variable  $i$  is declared to control the samples count, as well as a *stop* flag which is used to distinguish between inflation and deflation stages. Line 16 verifies whether the start button has been pressed, in that case, the mini electrovalve is closed and a loop from 0 to the maximum number of samples is began. Both during the inflation and deflation stages, the sensor reads the pressure of the cuff and stores the data on the global array *muestras*, on the next line the  $i$  value is increased. When the value of this variable is equal to 64, it is proceeded to calculate the autocorrelation of the samples, the value of  $i$  is reset, then it is tested if the last *dato* is less than or equal to 55 and if the *paro* signal is different from 1 (this means it is on the deflation stage). If this is positive the loop is broken and is followed by the calculation of the BP of the patient. In case  $i$  is different from

64, it is checked whether *dato* is less than 160 and *paro* is equal to 1 (this means it is on the inflation stage), hence the air pump is activated, otherwise the *paro* signal is set to 0 followed by the deactivation of the air pump, the mini electrovalve is opened to 11.69%, decreasing on a .80% the PWM work cycle every 20 mm Hg.

---

**Algorithm 1:** Procedimiento general para calcular PAS y PAD.

---

**Datos:** Máximo 3000 lecturas del sensor.  
**Resultado:** PAS y PAD del paciente.

**1 Variables Globales:**

```

2     indice ← 0
3     TOTAL_MUESTRAS ← 3000
4     N ← 64
5     muestras[N]
6     promedios[TOTAL_MUESTRAS/N]
7     valorMaxAuto[TOTAL_MUESTRAS/N]
8 begin
9     Inicialización:
10    Botones de la FPGA
11    Bus GPIO para controlar bomba de aire
12    /* 250 Khz con ciclo de trabajo al 50% */
13    Pulso PWM para controlar mini-electroválvula
14    while True do
15        i ← 0
16        paro ← 1
17        if botonInicio = 1 then
18            Cerrar mini-electroválvula
19            for j ← 0 to TOTAL_MUESTRAS do
20                leerSensor(dato)
21                muestras[i] ← dato
22                i ← i + 1
23                if i = N then
24                    calcularAutocorrelacion()
25                    i ← 0
26                    if dato ≤ 55 y paro ≠ 1 then
27                        break
28                /* Etapa de inflado del brazalete
29                */
30                if dato < 160 y paro = 1 then
31                    Activar bomba de aire
32                /* Etapa de desinflado del
33                brazalete */
34                else
35                    paro ← 0
36                    Desactivar bomba de aire
37                    Abrir mini-electroválvula al 11.69%
38                    Disminuir 0.80% al ciclo de trabajo de
39                    PWM cada 20 mm Hg
40            calcularPresionArterial()
41            paro ← 1

```

---

Fig. 5. General procedure for evaluate SBP and DBP.

---

```

1 Función: calcularAutocorrelacion()
3 /* Este arreglo contiene los 64 datos
   obtenidos por el sensor. */
5 Datos: muestras[N]
7 Resultado: Valor promedio de las 64 muestras y valor máximo
   de la Autocorrelación.
8 promedios[indice] ← Obtener el promedio
10 /* Calcular la Autocorrelación de las 64
   muestras. */
11 b ← 0
12 for m ← 0 to N do
13   rx ← 0
14   for n ← 0 to N - m do
15     rx ← rx + ((muestras[n] - promedios[indice]) *
16                (muestras[n + m] - promedios[indice]))
17   resultado[b] ← rx
18   b ← b + 1
19 valorMaxAuto[indice] ← Valor máximo de resultado[N]
20 indice ← indice + 1

```

---

Fig. 6. Autocorrelation function.

## 4 Data Display Software

The data display software consists of a test web program where the medical personnel in charge of reading the BP measurements of the patients can register. Said personnel can register new patients on the database using the same web site, as well as generate reports about the performed measurements. The use case for the medical personnel is shown on Fig. 7 and the main screen of the system is shown on Fig. 8.

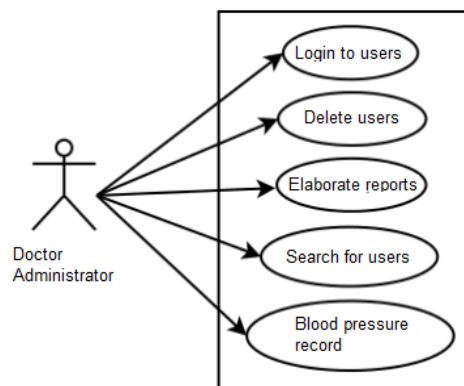


Fig. 7. Use cases for medical staff.

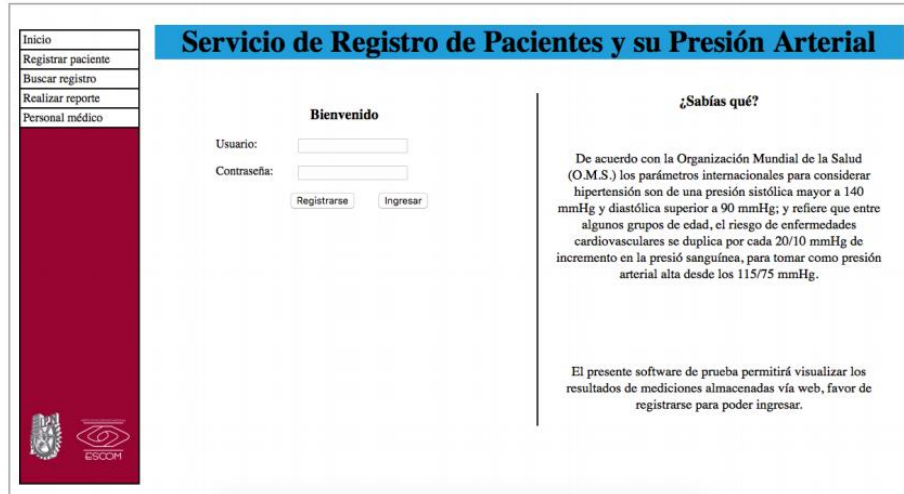


Fig. 8. Main system screen.

## 5 Testing and Results

The hardware configuration of the embedded system was done using *Xilinx Platform Studio (XPS)*. Firstly, the required controllers for the FPGA Nexys 3 were selected, the number of processors to be used, as well as the Soft-core Microblaze and the GPIO, UART, Timer/Counter and I2C cores. Secondly, space on memory was assigned, then the input/output signals were established linking every one of them to the FPGA pins. Finally, the project was compiled to generate a bitstream file. The embedded system is shown on Fig. 9.



Fig. 9. Final prototype of the embedded system.

The software configuration was done using Xilinx Software Development Kit (SDK). In first place, it is necessary to import the bitstream file to create the support files for each core. Then a C language project was created where a program which implements the previously mentioned pseudocode was created. After that, the project

was compiled so as to create a *download bit* output file. This is the file which is burned on the FPGA y grants functionality to the embedded system.

Considering that the method used is the oscillometric one, the deflation stage is of vital importance given that it is during this stage where the pressure oscillations present on the cuff, therefor deflation test were done to the mini electrovalve, varying the PWM pulse frequency of the FPGA as well as its work cycle, being able to prove that with a 250 KHz frequency and a 11.69% work cycle the air was expelled properly from the cuff but only on the 160 to 120 mm Hg range. This is because the pressure on the cuff decreases gradually however the mini electrovalve continues letting the air flow at a steadfast rate. Therefore, the work cycle was reduced by a 0.8064% every time the pressure on the cuff decreased by 20 mmHg beginning at 120 mmHg given that the mini electrovalve is activated on low.

**Table 1.** Measurement results for 25 people.

No.	Valor del Sistema (VdS) (PAS / PAD) (mm Hg)	Valor del Baumanómetro (VdB)(PAS / PAD) (mm Hg)	Error Absoluto ( VdS-VdB ) (mm Hg)	Error Relativo ( VdS-VdB /VdS)
1	115/ 67	120 / 68	5/1	4.3 % / 1.49 %
2	123/ 73	123/ 69	0/4	0 % / 5.47 %
3	137/ 73	129/ 73	8/0	5.83 % / 0 %
4	117/ 71	122/ 69	5/2	4.27 % / 2.8 %
5	117/ 70	117/ 66	0/4	0 % / 5.71 %
6	130/ 65	137/ 77	7/12	5.38 % / 18.46 %
7	125/ 80	119/ 67	6/13	4.8 % / 16.25 %
8	137/ 79	137/ 77	0/2	0 % / 2.5 %
9	113/ 65	106/ 60	7/5	6.19 % / 7.69 %
10	100/ 60	100/ 62	0/2	0 % / 3.33 %
11	112/ 78	116/ 72	4/6	3.57 % / 7.69 %
12	127/ 84	120/ 68	7/16	5.83 % / 19.04 %
13	129/ 81	124/ 78	5/3	3.87 % / 3.7 %
14	114/ 75	115/ 77	1/2	0.87 % / 2.66 %
15	122/ 80	119/ 74	3/6	2.45 % / 7.5 %
16	124/ 68	125/ 78	1/10	0.8 % / 14.7 %
17	119/ 83	118/ 74	1/11	0.84 % / 13.25 %
18	115/ 65	119/ 81	4/16	3.47 % / 24.61 %
19	105/ 68	102/ 69	3/1	2.85 % / 1.47 %
20	111/ 67	111/ 75	0/8	0 % / 11.94 %
21	137/ 80	136/ 77	1/3	0.72 % / 3.75 %
22	111/ 81	111/ 69	0/12	0 % / 14.81 %
23	123/ 67	120/ 75	3/8	2.43 % / 11.94 %
24	122/ 80	119/ 74	3/6	2.45 % / 7.5 %
25	111/ 73	117/ 73	6/0	5.4 % / 0 %

Also, a full opening signal was set when the cuff presented a pressure less than or equal to 55 mm Hg because after over 50 test we observed that if the PWM was left until the cuff deflated completely, the SBP and DBP calculation was not done correctly. Likewise, even though in theory the cuff has to be deflated at a 3 mm Hg/s rate in order to obtain a correct blood pressure measurement, in practice it was observed that the

OMRON commercial baumanometer model HEM-742INT, which was used for comparisons, deflated the cuff at a 5 mm Hg/s rate. On this embedded system, the deflation stage was done within a range of 20 to 25 seconds at a 4.5 mm Hg/s rate.

There were made 25 measurements to the same amount of people with the OMRON electronic baumanometer and this embedded system. This baumanometer was taken as a reference value because the WHO and the Panamerican Health Organization (OPS, given its Spanish acronym) establish that well-kept and properly calibrated and validated mercury free tensiometers or baumanometers provide an accuracy commensurable to non-mercury free devices [12], in this case the OMRON baumanometers are distinguished for following strict calibration protocols and for sticking to several international standards (ANSI and ISO), this model in particular is validated by the British Hypertension Society and has an accuracy of  $\pm 3$  mm Hg [13]. The results are shown on Table I. The absolute and relative errors of each measurement were calculated, the SBP presented a variation between 0 and 8 mm Hg which corresponds to 0 and 6.2% relative error and the mean SBP relative error was 2.65%. On the other side, the DBP presented a wide absolute error margin going from 0 to 16 mm Hg which correlates to relative errors between 0 and 24.7%, with a 8.39% mean relative error.

## **6 Conclusions**

Despite being on the 21<sup>st</sup> century, Mexico displays severe lateness on health care, mainly as a consequence of overpopulation and lack of medical personnel to provide satisfyingly the human resources requirements of this area, without mentioning that in most administrative and auscultation processes which are carried out manually, the presence of medical personal is needed to perform it. In the search of reducing the lack of health care problem in the social context, arises the need for the development of new methodologies that incorporate technology.

The present work was able to design and implement an embedded system on a FPGA to monitoring the blood pressure of a patient, automatizing the data acquisition from the digital sensor controlling an air pump for the cuff inflation stage and a mini electrovalve controlled via a PWM signal which allows for regulating the cuff's deflation rate. The values of the work cycle and frequency were obtained in an experimental way.

An algorithm was proposed to calculate the blood pressure on the embedded system taking the oscillometric method as foundation and performing the autocorrelation of the samples taken by the sensor. Following the same line, a program was designed and implemented to allow the storage of doctors and patients register information, as well as the blood pressure data received from the FPGA and generate reports on this.

By using reconfigurable technology, we can customize the architecture to fit our needs. The proposed system allows an easy addition of other vital signs to the architecture, along with a different amount of them on contrast to the systems that use microcontrollers, DSP or processors (ARM, MIPS, etc) with a predefined architecture.

**Acknowledgment:** The authors would like to thank the Graduate and Research Division of the National Polytechnic Institute who contributed to the development of this work

through the SIP 20161893 project. Also to the participation of Jesus Gutierrez Sanchez, P. Eng and Oscar Ramirez Garcia, P. Eng on this project.

## References

1. Ying-Wen Bai *et al.*: Design and Implementation of an Embedded Monitor System for Body Breath Detection by Using Image Processing Methods. In: Digest of Technical Papers International Conference on Consumer Electronics (ICCE), pp. 193–194 (2010)
2. Siddharth, N., Sasikala, M.: Design of Vital Sign Monitor based on Wireless Sensor Networks and Telemedicine Technology. In: International Conference Green Computing Communication and Electrical Engineering (ICGCCEE), pp 1–5 (2014)
3. Baccini, P.H. *et al.*: Developing an Affective Point-of-Care Technology. In: IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE), pp. 77 – 84 (2014)
4. Huang, Han-Pang, Hsu, Lu-Pei: Development of a Wearable Biomedical Health-Care System. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1760–1765 (2005)
5. Kalovrektis, K. *et al.*: Development of Wireless embedded system using ZigBEE protocol to avoid white Gaussian noise and 50 Hz Power line noise in ECG and Pressure Blood signals. In: Panhellenic Conference on Informatics PCI '08, pp. 113–117 (2008)
6. Rozman, C., Cardellach, F.: Medicina Interna. XVII Edición, Volumen I, Capítulo 60 (2012)
7. O'Brien, E., Asmar, R., Beilin, L., Inai, Y., Mancia, G.: European Society of Hipertension: Recommendations for conventional, ambulatory and home blood pressure measurement, J. Hypertens 21, 821-48 (2003)
8. Honeywell, Inc.: HSCDAND015PG2A3 Data sheet (2016)
9. Xilinx, Inc.: Spartan6 – Family overview (2011)
10. Digilent, Inc.: Nexys 3 Board Reference Manual (2015)
11. Babbs, C.F.: Oscillometric measurement of systolic and diastolic blood pressures validated in a physiologic mathematical model (2012)
12. OMS y OPS: Reemplazo de los termómetros y de los tensiómetros de mercurio en la atención de salud. Guía técnica (2013)
13. Destro, D.G., Araujo, R., Mendes, A., Alves, M., Doederlein, M., de Oliveira, A.: Validación del monitor de medición de la presión arterial Omron HEM 742. Arq. Bras. Cardiol 92(1) (2009)

# Cryptography in Wireless Network Penetration Testing

Claudio Casado<sup>1</sup>, Cristian Barría<sup>2</sup>, Lorena Galeazzi<sup>3</sup>

<sup>1</sup> Universidad Mayor, Santiago, Chile

<sup>2</sup> Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

<sup>3</sup> Instituto Profesional DuocUC, Santiago, Chile

<sup>1</sup> claudio.casado1@mayor.cl, <sup>2</sup> cristian.barría@ucv.cl

<sup>3</sup> l.gale.ava@gmail.com

**Abstract.** At present, wireless accesses are experiencing exponential growth worldwide, not being unrelated to various attacks on the integrity, availability and confidentiality of information; Resulting in the implementation of security measures. Based on the above, the protection of the data is of transcendental importance and together with this the implementation of cryptographic systems at the level of organizations and end users. This is why the present research exposes the different variables that must be considered during security tests to obtain information, and with the purpose of achieving a contribution to protection methodologies that are incorporated in the different components of the wireless networks.

**Keywords:** wireless, pentesting, criptography.

## 1 Introduction

Based on the internet service exposure, this has been a key role in the technological growth of telecommunications, increasing its development curve, thanks to the ability to entertain the user to maintain an uninterrupted connection, where society demands the creating new systems that satisfy their demand and needs, directing technological progress towards the transfer of information independent of physical media such as cable. The foregoing is reflected in figures for wireless technology growth at the national level, where in recent years internet access points have grown by 45.3% between 2014 and 2016, reaching 73.8% of the population, playing at the same time in a more important activity in the activity of the common things, to the economic transactions and the private and public communications [2].

In spite of the above, the suppliers of the item have focused their efforts on improvements and advances oriented to functionalities and quality of the connections, without considering the security that should have the devices that are part of the network [3].

Along with the growth of the above-mentioned technology, there is a similar growth in the associated security incidents. Under an international context, between 2016 and 2017, the concerns of companies in Latin America on information security issues have increased by an average of four percentage points [4].

On the other hand, at the level of end users, the gap between them and the security that revolves around the technology in question reaches worrying levels, since, according to the company Norton, only 48% of users are able to determine if a wireless network is secure. Within the same study, 35% of users have at least one unprotected device connected to their network [5]. Cyber-attacks and espionage activities in the network, massive interception of telecommunications networks, disruption of the internet service, espionage and attacks against critical infrastructures and governmental entities, have set the guidelines in this area [2].

In view of the above, the existence of studies that support Information Security Management Systems (ISMS) with accurate data to obtain a baseline assessment of the risk associated with the null or inadequate implementation of the different methods is problematic. cryptographic devices that provide security for wireless networks. As a consequence, it is sought, as a contribution, to propose a risk assessment associated with the aforementioned variable that will serve as a basis to support both security testing and the preparation of security plans in support of ISMS. The second section presents related works that explain the current encryption mechanisms and the variables that could imply a risk for the network. In the third section, a risk analysis based on OWASP methodology is carried out. Finally, in the fourth section, we present the conclusions, analyzes and future work that emerge from the results of the risk analysis.

## **2 Related Work**

The main characteristic of wireless networks is the ability to generate a connection between a transmitting device and a receiving device, taking into account the only condition in which it must be within the transmission range of the signal remittent. The independence based on its wireless structure provides the end users mobility, faculty responsible for the momentum in the vertiginous growth of this technology [6].

Given the above, one of the biggest problems to be considered is security in Wi-Fi networks where Detect, isolate and prevent malicious acts that, if realized, would cause much greater effects on the network is the goal. It is thus that the most important actions to take into account to safeguard the security of a network and its information, are the identification and authentication of users in addition to the classification of the risks associated with these characteristics [7].

In response to the need to provide user identification and authentication to wireless networks, the Institute of Electrical and Electronic Engineers (IEEE) implements the 802.11 standard in 1999, which registers remarkable advances as time passes and technology advances. only in the frequencies that it uses and its maximum speed of transmission, but in the mechanisms of encryption that it uses to grant security to the moment a user wants to enter the network [6].

Some of the most important security mechanisms involve encryption and key management mechanisms to achieve the aforementioned security features [8] [3].

The cryptographic mechanisms of the IEEE standard mentioned above are: WEP, WPA and WPA2, in chronological order of implementation. The former proved, through numerous investigations, to be extremely vulnerable as it uses RC4 encryption, with public key transmitted in clear text and does not offer end-to-end security.

The second is more secure than WEP because it was created as a corrective measure within the 802.11i update of the IEEE standard given the insecurity presented by its predecessor. WPA includes integrity control to the messages it manages. It uses Michael encryption algorithm (MIC) and TKIP (Temporal Key Integrity Protocol). Finally, WPA2 was created to convince its predecessor as a secure encryption mechanism, being more robust, efficient and complex to break than WEP and WPA since in addition to the features incorporated in the latter, it includes 4 Ways Handshake and CBC-MAC encryption algorithm, in addition to CCMP-AES as security protocol [8, 3]. As a means of authentication, enterprise networks use the EAP framework (with different versions of it, such as LEAP, PEAP, EAP-TTLS, EAP-TLS, EAP-FAST, EAP-MD5) Pre-Shared Key (PSK). Both features are present in both WPA and WPA2. The encryption mechanisms described above are still in force today, but these are the ones that predominate in the default configurations in WEP, which, despite their validity as a crypto-do mechanism, does not guarantee the minimum of security that justifies its activation by default, making it tend to be considered obsolete. Details on the operation and characteristics of the above are summarized in Table 1 [8].

**Table 1.** Encryption protocols in wireless networks and their characteristics.

Characteristics	WEP	802.11i	
		WPA	WPA2
Security Protocol	RC4	TKIP	CCMP
Cipher	RC4	RC4	AES
Key Length	40 or 104 bits	128 bits encryption 64 bits authentication	128 bits
Key Life	24 bit IV	48 bit IV	
Key Generation	Concatenation	Two phase mixing function	Not needed
Data Integrity	CRC-32	Michael	CBC-MAC
Header Integrity	None	Michael	CBC-MAC
Replay Protection	None	Packet Number	
Key Management	None	EAS-based	
Authentication	Open or shared key	802.11x or PSK	
Wi-Fi Alliance Certificate (WPS)	Active/No Active		

Although WPS is not a feature of encryption mechanisms, it affects the security they seek to provide, since in their desire to facilitate the installation and configuring of known devices on the network, it leaves aside, without compensation, the security provided by the encryption techniques associated with authentication [9].

Once the characteristics associated to the different cipher-do mechanisms have been described, a more specific description of the variables that represent, today, a security risk given by these mechanisms (see Table 2).

**Table 2.** Definition of variables.

	Variable	Description
WEP	Key	Corresponds to the key of access to the network defined by the administrator of this one [3]
	Keystream	Corresponds to the result of the XOR binary operation between a WEP key and a given Initialization (IV) Vector [6]
WPA	PSK	It corresponds to the authentication process for a specific client or device. It is known as four-way handshake since it uses 4 validations of the authentication [3]
	EAP-Handshake	It works in conjunction with WPA-PSK to add a secure "wrapper" to the information that travels between the AP and the client when performing a four-way handshake [3, 8].
	TKIP Encryption	Encryption type that allows to deliver confidentiality to data packets traveling through the network [7]
WPA2	4-way Handshake	Authentication process used by this protocol [10].
WPS	PIN	Functionality present in APs using the eight-digit PIN exchange with client device for easy connection, installation and configuration [9].

### 3 Risk Analysis

As defined by the Open Web Application Security Project (OWASP) in its risk analysis methodology, it is calculated according to the following equation [11]:

$$Risk = Likelihood * Impact. \quad (1)$$

The first step to carry out this methodology is to identify the variables that could represent a risk. Secondly, an estimate must be made of the probability that exploitation of such risk will be effective based on the identified vulnerabilities. Finally, one must determine the impact that the exploitation of each of the vulnerabilities can have. To analyze qualitative variables OWASP recommends the assignment of values between one and ten that allow to establish ranges that define if a vulnerability and / or associated risk correspond to High, Medium or Low. The methodology in question recommends the use of sub-factors to establish a more accurate quantitative assessment based on qualitative variables [11].

For the calculation of the risk (R) associated to a variable, in the present work it is considered the use of two sub-factors for the calculation of the probability of occurrence. These are the popularity of the attack (P) and its simplicity (S) where its average multiplied by the impact factor (I), results in the desired estimate. Given the above, the equation for risk analysis would be the following:

$$R = \left( \frac{P * S}{2} \right) * I. \quad (2)$$

Given the above (equation 2), the measurement levels for probability and impact are defined by associating at low level the values from one to less than four, as mean scores between one and less than seven, and finally as high ones that go from seven to ten. Given the variables and methodology defined above, each of these is valued for

subsequent risk analysis. A risk weighting will be performed for each of the malicious acts, extracting in turn an overall risk average pertaining to each of the cryptographic variables reviewed in this work: WEP, WPA / WPA2, WPS [3, 12]. Table 3 shows the results of the risk analysis carried out for each type of attack.

**Table 3.** Risk weighting for attacks associated with cryptographic variables.

Protocol	Variable	Attack	P	S	I
WEP	WEP Key Recovery	FiOS SSID WEP Key Recovery	9	10	8
		Nessus Datacom 21-bit attack	8	9	8
		Dictionary Attack	4	10	8
		Cryptographic Attack	7	5	8
		Using Aircrack-ng (Client Attached)	7	5	8
	Keystream Recovery	Chop-Chop Attack	4	4	7
WPA	Breaking Authentication WPA-PSK	Obtaining the Four-Way Handshake	7	4	9
		Cracking the PSK	7	4	9
		Decrypting WPA-PSK Captures	6	4	6
	Obtaining the EAP-Handshake	Attacking LEAP	4	6	8
		Attacking PEAP & EAP-TTLS	7	4	9
		Attacking EAP-TLS	1	1	10
		Attacking EAP-FAST	5	5	9
		Attacking EAP-MD5	4	7	7
	Braking Encryption TKIP	Beck-Tews Attack	4	4	8
WPA2	4-way Handshake	KRAK Attack	3	2	7
WPS	WPS Attack	Brute-force Hack	10	10	10

Once the previous tabulation has been performed, the calculation of the weighted general risk is performed for each of the encryption methodologies, as shown in Table 4.

**Table 4.** Likelihood and Impact Weighting for WEP, WPA, WPA2 and WPS.

WEP			WPA			WPA2			WPS		
Likelihood		Impact	Likelihood		Impact	Likelihood		Impact	Likelihood		Impact
P	S	I	P	S	I	P	S	I	P	S	I
6,5	7,2	7,8	5	4,3	8,3	3	2	7	10	10	10
6,8		7,8	4,7		8,3	2,5		7	10		10

After the previous tabulation was carried out, the calculation of the general risk weighted for each cryptographic variable according to the methodology proposed by OWASP (Table 5).

**Table 5.** Matrix for obtaining real risk for each cryptographic variable.

		Overall Risk Severity		
Impact	High	Medium (WPA2)	High (WEP y WPA)	Critical (WPS)
	Medium	Low	Medium	High
	Low	Note	Low	Medium
		Low	Medium	High
	Likelihood			

## 4 Conclusions and Future Work

It can be concluded from the present work that although cryptography and encryption mechanisms seek to grant users authentication in a network, they will not fully comply with this role if the security configurations necessary to reduce the risk present in this type of technologies. Factors such as the obsolescence of some of these mechanisms, make a network insecure despite having the technology and corresponding updates to achieve this goal. The deactivation of WPS and the use of WPA2 within the configurations of the router are good basic and recommended minimum practices to mitigate the present risk in a network.

Although a risk analysis applied in an organization considers more sub-factors of analysis specific to the business context, through the present work we seek to contribute with a risk analysis that serves as a basis for the generation of con-science about the null or incorrect application of security on a network infrastructure. This allows both to support strategies and methodologies for performing security tests on this type of technology, and to support ISMSs through clear information that serves as a basis for the development of plans and policies for associated good safety practices to the protection of wireless networks.

As future work, it is proposed to carry out similar work to quantify the risk associated with other variables present in the wireless networks, further reinforcing the contribution to the problem raised in this research, supporting both the ISMS and the associated pentesting methodologies.

## References

1. Salvetti, D.: Redes Wireless. 1era Edición, Buenos Aires: Fox Andina, Dalaga (2011)
2. National Security Council of Chile: National Cybersecurity Policy (NSP) for 2017–2022. <https://www.ciberseguridad.gob.cl/media/2017/05/NCSP-ENG.pdf>
3. Cache, J., Wright, J., Liu, V.: Hacking Wireless Exposed: Wireless Security Secrets & Solutions. McGraw Hill (2010)
4. Laboratorios ESET: ESET Security Report Latinoamérica (2017)
5. Symantec Corporation: Norton Cyber Security Insights Report (2016)

6. Wadhwa, U.: Wireless Network Security: Tough Times. In: International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 1022–1025 (2015)
7. Troya, A., Astudillo, J., Romero, C., Sáenz, F., Díaz, J.: Vulnerability Detection in 802.11i Wireless Networks Through Link Layer Analysis. In: IEEE Latin-America Conference on Communications (LATINCOM) (2014)
8. Hassan Adnan, A., Abdirazak, M., Shamsuzzaman Said, A., Anam, T., Zaman Khan, S., Mahmudur Rahman, M., Musse Omar, M.: A comparative study of WLAN security protocols: WPA, WPA2. In: International Conference on Advances in Electrical Engineering, N° 3, pp. 165–169 (2015)
9. Instituto Nacional de Ciberseguridad de España: Qué es WPS Pin y por qué debes desactivarlo. Oficina de Seguridad del Internauta (2014)
10. Vanhoef, M., Piessens, F.: Predicting, Decrypting, and Abusing WPA2/802.11 Group Keys. USENIX, vol. XXV, pp. 673–688 (2017)
11. OWASP Foundation: About The Open Web Application Security Project (2018)
12. Gonzalez-Pérez, P., Sanchez-Garcés, G., Soriano de la Cámara, J.M.: Pentesting con Kali. OXWord (2013)



# Computer Vision Algorithm Implementing Geometric Morphometry to the Shape Analysis

L. Jhonathan Flores-Guarneros, B. Esther Carvajal-Gómez

Instituto Politécnico Nacional, Escuela Superior de Cómputo,  
Sección de Estudios de Posgrado e Investigación, Mexico City, Mexico  
lfloresg1507@alumno.ipn.mx, becarvajal@ipn.mx

**Abstract.** The present work defines the guidelines to develop a shape analyzing system in order to classify an image set. The classification process considered geometric morphometry as principal theory for pattern recognition and to extract scale-rotation-invariant key points. This research also compares some theories and techniques for pre-processing and segmentation steps of the general image identification process due to sample images. The research's output considers an automated system to sex classify sea turtle hatchlings according to the shell morphology patterns.

**Keywords:** computer vision, digital image processing, geometric morphometry.

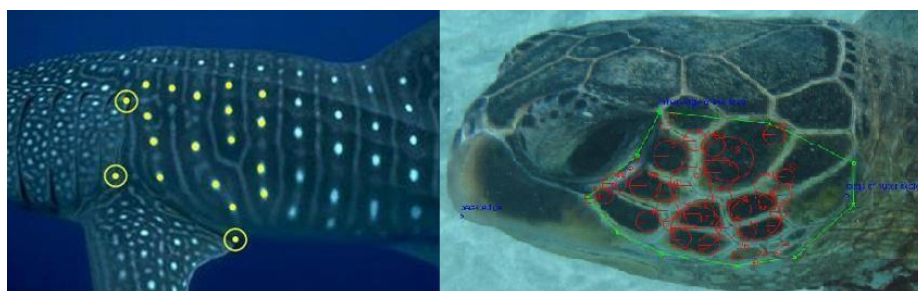
## 1 Introduction

Computer vision is a science focused on the development of automatic systems to simulate the biological visual perception system. The principal purpose of this science is to allow machines to make decisions recognizing and identifying objects around them. The computer vision systems are based on algorithms to analyze information extracted from digital images. These algorithms can obtain information about physical and structural characteristics such as the shape, the color and some metrics [1-5].

Computer vision systems are used in many different sciences to help scientists to make and to support decisions. The results of a shape analysis could diagnose a disease, confirm the identity of a person or classify a group of products. In Biology, the implementation of these systems aims to assist biologists in some identification processes. Particularly in this field, it is needed to get and to analyze some shape patterns to recognize an individual [6].

In 2014, Navarro [7] described the relationship between sea turtle hatching statistics and climate phenomena in nesting lands. He also suggested the difference between male-female sea turtle hatchlings as principal indicator to predict global warming progress. Although, the method to get exactly the genre of a sea turtle implies to kill the hatchling, some biologists pose new techniques to infer this type of information by analyzing specific shape patterns in the shell. These techniques consider that the morphological structure of each individual is unique so that is possible to classify animals studying the biological structure to find classification patterns [8,9]. Some other authors explain the necessity of applying quantitative methods (based on metrics)

to get accurate results. Geometric morphometry as a quantitative theory for shape analysis shows the benefits of its implementation in the general image identification process [9-11].



**Fig. 1.** Examples of patterns based on spots dispersion (left) and shape (right) to identify an individual. Obtained from [12] and [13] respectively.

These days, there is software created to assist and to make decisions based on quantitative characteristics. Avoiding subjective criteria in biologic shape analysis is real important to get accurate classifications. The Interactive Individual Identification System ( $I^3S$ ) generates a unique footprint for each specimen analyzed as shown in Fig. 1. The  $I^3S$  Pattern version automatically gets a recognition pattern based on the key points of the region of interest. The major disadvantage of this tool is that the final results depend on personal criteria of the scientist that is doing the analysis [12,13]. C. Town *et al.* [14] posed the “Manta Matcher” as an automated system to identify manta rays by analyzing key-point features extracted from images. The system was innovative due to the implementation of an algorithm with adaptive phases to reduce image noise and to accurate contrast equalization. However, the restrictions are the same; biologists need to supervise all the process and to make decisions for example the user must assure that the images are in the selected normal position.

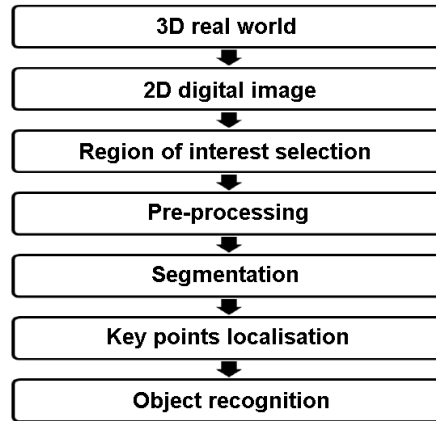
Thus the present work defines the guidelines to develop a shape analyzing system to get a correct pattern recognition and classification of sea turtle hatchlings into two groups (male or female) by applying geometric morphometry. The system will implement computer vision algorithms to process 2D digital images of sea turtle hatchlings and to approach the aim.

## 2 Methodology

An abstraction of the general image processing methodology could be appreciated in the diagram of Fig. 2. This methodology defined the order in which different algorithms should be implemented for the recognition process. For this reason, it could be used to guide the research approach presented in this paper.

Third stage: Region of Interest (ROI) selection is really important and refers to select the region of the individual where are enough landmarks (key points) located to create a unique shape pattern. The next phase considers the implementation of pre- processing algorithms. Sometimes it is needed the noise removing applying filtering algorithms to

vanish and homogenize the information of the region. There are other many times that the emphasis of existing differences in the information is needed as well, so a contrast equalization algorithm is applied.



**Fig. 2.** General image-processing methodology.

The fifth stage refers to separate the image information into different sets. This could be considered as the first approach to the object recognition because the structure of each figure in the image is defined in accordance with those sets. The next step is to locate the information used to create the shape patterns. These patterns could be described by some important points (landmarks in Biology) or the figure edges. The final phase is related to the recognition process. Once the landmarks have been located, a recognition algorithm is trained to recognize the patterns and classify the images into different groups.

### 2.1 Sample Images

As said before the principal aim of this work is to verify the results of the object classification based on shape analysis by applying geometric morphometry. To accomplish that aim two sets of images were selected, but the only condition to apply this technique is to have shape patterns for object recognition. The first image set considers around 30 images of turtle hatchlings for a sex classification system and the second includes images of basic geometric figures (squares and rectangles) with controlled conditions about color intensity. The second set only had the purpose to validate the application of geometric morphometry for pattern classification based on any shape into groups.

## 3 Results

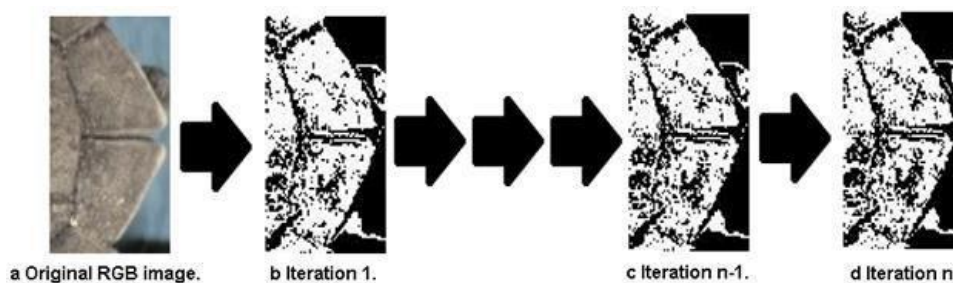
For the turtle hatchling images the region of interest was selected according to biologist recommendations as shown in Fig. 3. The type of information contained in each image of this set needed to be treated applying pre-processing algorithms. In Fig. 3, the result images of the implementation of these algorithms are shown.



**Fig. 3.** Preprocessing image results.

In the segmentation stage different algorithms were implemented: threshold, statistical analysis and fuzzy clustering. Fuzzy clustering algorithm analyzes and classifies complex parameters based on fuzzy logic fundamentals [15].

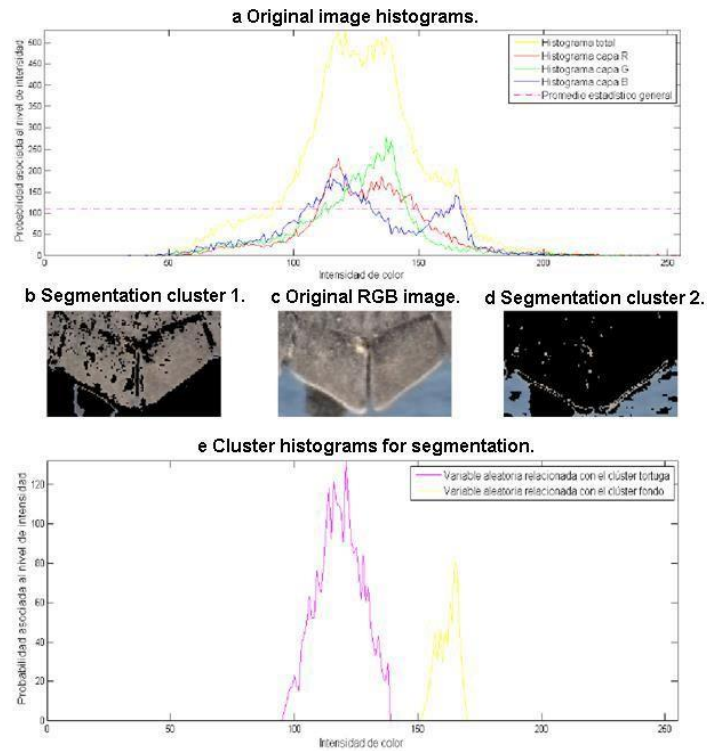
Fig. 4 shows the result of fuzzy clustering segmentation, as an iterative algorithm could take long time to get the accurate classification clusters. Fig. 5 shows the alternative method for image segmentation by means of the image statistical analysis. This algorithm applies the principal statistical moments.



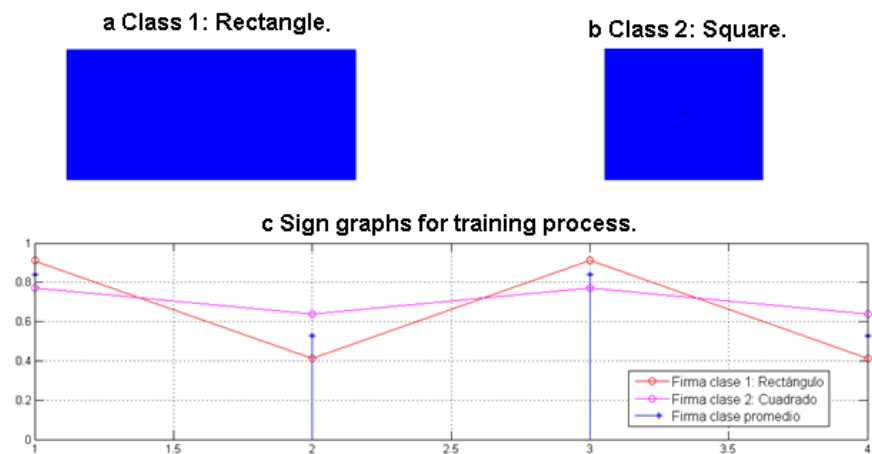
**Fig. 4.** Results of fuzzy clustering segmentation process.

The set on geometric figure images was used to validate the implementation of the geometric morphometry algorithm. Fig. 6 shows the geometric figures used to train the classification algorithm.

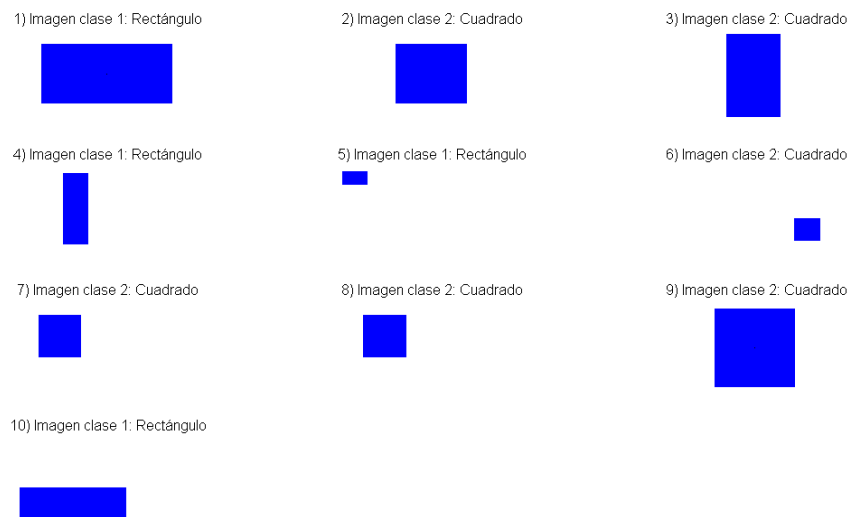
The results of classification process implementing the geometric morphometry theory are shown in Fig. 7. The geometric figure image set was integrated by ten different images. The obtained results applying this theory are also in Fig. 7.



**Fig. 5.** Segmentation process considering statistical image analysis.



**Fig. 6.** Income information for classification process.



**Fig. 7.** Results and images of classification process.

## 4 Conclusions

This paper only presents results until the fifth stage of the general image processing methodology applied to the sex classification system. The third phase results to be really important as [7,8] and [9] mentioned before. The selection of the right region of interest could make easier the activities of the following steps.

Those previous works also mentioned the necessity of validate with the specialist all the results automatically generated. This condition could be explained with the results obtained with different segmentation algorithms.

Segmentation results are considered as not good enough to implement any of the automatic algorithms studied. In conclusion in this part of the process could be better consider the opinion of the biologist. The automatic system will only help to make easier the localization of key points, but the biologist will select where this points are.

The shape analyzing algorithm has been validated with the geometric figure image set, so it could be implemented in the pattern recognition system to identify the sex of a turtle hatchling. This work is still in progress.

The next step is to continue implementing the shape analyzing algorithm to the recognition system and in a long term could be to implement this new assistant system in a mobile device to give facilities and help the turtle protection.

## References

1. Trembley, J.P., Manohar, R.: *Visión Computacional. Matemáticas discretas con aplicación a las ciencias de la computación*, México: CECSA (2000)
2. Pajares, G., de la Cruz, J. M.: *Visión artificial. Visión por computador*, 1ra. ed. México: Alfaomega, pp. 1–12 (2002)

3. Cuevas, E., *et al.*: Procesamiento digital de imágenes con MATLAB y Simulink, 1ra. ed. México: Alfaomega (2010)
4. Schenk, T.F.: Procesamiento Digital de imágenes. Fotogrametría digital. Catalunya, Marcombo, vol. 1, ch. 3, pp. 66–96 (2002)
5. Sossa-Azuela, J.H.: Introducción. In: Visión artificial: rasgos descriptores para el reconocimiento de objetos, 1ra. ed. Madrid: RaMa, ch. 1, pp. 19–26 (2013)
6. Pajares, G., de la Cruz, J.M.: Aplicaciones de la visión artificial. Visión por computador: Imágenes digitales y aplicaciones, 3ra. ed. México: RaMa, pp. 4–6 (2008)
7. Navarro, E.J.: Efecto de la temperatura de incubación y la diferenciación sexual sobre la morfología de crías de tortuga Golfina *Leopidochelys olivácea*. M.S. thesis, Dept. Bio. Marina, UNAM, Mazatlán, México (2014)
8. Franco-Moreno, R.A.: Morfología y desempeño del aparato mandibular de seis especies- de peces Ictiófagos demersales, asociados a los fondos blandos de la plataforma continental de Nayarit, Sinaloa. M.S. thesis, CICIMAR, IPN, La Paz, México (2011)
9. Toro, M.V., Manriquez, G., Suazo, I.: Morfometría geométrica y el estudio de las formas biológicas: de la morfología descriptiva a la morfología cuantitativa. *Int. J. Morphol.* 28(4), 977–990 (2010)
10. López-Galán, A.: Morfometría geométrica: el estudio de la forma y su aplicación en biología. *Temas de Ciencia y Tecnología* 19(55), 53–59 (2015)
11. Mantini, S., Ripani, M.: Modern morphometry: new perspectives in physical anthropology. *New Biotechnology*, Elsevier 25(5), 325–330 (2009)
12. *I3S pattern manual version 4.0.2*, Free software foundation, Inc., 59 Temple Place – Suite 330, Boston (2014)
13. Speed, C.W., *et al.*: Spot the match – wildlife photo-identification using information theory. *Frontiers in Zoology* 4(2) (2007)
14. Town, C., *et al.*: Manta Matcher: Automated photographic identification of manta rays using keypoint features. *Ecology and Evolution* 3(7) (2013)
15. Cano-Plata, E., Cano-Plata, C.: Análisis de los algoritmos de agrupamiento borroso para detectar asimetría de información. *Economía y Administración*, No. 70 (2008)



# Application of the Shamir Threshold Scheme to a System for Safely Storing and Sharing Experimental Clinical Studies in Accordance with the Official Mexican Standard NOM-024-SSA3-2012

José Daniel Pérez Ramírez<sup>1</sup>, Lorena Chávez Nava Olguín<sup>1</sup>, Blanca Alicia Rico Jiménez<sup>1</sup>,  
Carlos Hernández Nava<sup>1</sup>, Laura Ivoone Garay Jiménez<sup>2</sup>

<sup>1</sup> Instituto Politécnico Nacional, UPIITA, Mexico City, Mexico  
{danprjs, lorecaol11, hernandez.nava}@gmail.com  
bricoj@ipn.mx

<sup>2</sup> Instituto Politécnico Nacional, UPIITA, SEPI, Mexico City, Mexico  
lgaray@ipn.mx

**Abstract:** This article presents an application of the Shamir threshold scheme to safeguard the confidentiality of experimental clinical studies that are stored, managed, and shared among different users to comply with the provisions for their safety in the Official Mexican Standard NOM -024-SSA3-2012. Because these types of databases contain section with sensitive information, a mechanism must be in place to ensure the safety and confidentiality of experimental clinical trials at different levels. In this case, it is important to restrict access to confidential information to only certain users. One solution is to use a symmetric encryption scheme, but the exchange of the encryption key without undermining system security could be a problem, because users are not always in the same place and time. Using the Shamir schema with a minimum threshold of two, dividing the key into as many subkeys as users exist in the system and assigning one to each user system, it will be possible for them to know the complete secret key when combining with another one to be able to access the confidential information, once the user had been identified and validated by the system. The preliminary results have shown that this solution is viable and allows to comply with the safety requirements established by the standard.

**Keywords:** Shamir threshold scheme, sensitive information, information sharing, sharing of secrets, cryptography.

## 1 Introduction

Nowadays, there are research laboratories and programs that offer the opportunity to clinical staff, researchers, and students from different knowledge background to collect medical information through experimental clinical studies to be shared and analyzed by multidisciplinary groups of physicians, telematics, biomedical and bionics engineers to search for new knowledge and perform an effective collaboration between groups of scientists [1] [2]. Besides the collected information grows exponentially in institutions such as IMSS, ISSSTE and SEDESA, which are migrating their information to digital files that include different types of files, ranging from images, clinical information,

questionnaires, even images of hand written notes. On the other hand, part of this information, according to Official Mexican Standard NOM-024-SSA3-2012 [3], could be sensitive and/or confidential information, so only personnel with certain privileges should have access to it, so a tool is needed to help to share such information without impairing its security and confidentiality.

According to the Standard, sensitive information is all information that contains the minimum data for the identification of persons, such as: CURP, first name, last name, name, date of birth, state of birth, sex, among others. [3] However, many of the interesting clinical conditions of people require partially some of this sensitive information in order to establish whether the obtained values are congruent or not with normal or healthy clinical status. Another classic example is epidemiological studies where the events are frequently associated with the area where the patient is currently living, where he was born, what type of activities he performs, income, etc. [4]

Considering that into research laboratories where there is a staff rotation and several people could participate in the research then it is not always the same person who performs the collection of clinical information to which prepares the databases or analyzes it, and generates new information or creates the theoretical models. Then a management tool is necessary to allow the exchange in a secure and closed collaborating network and that the shared access to the generated information could be performed without prejudice to its security and confidentiality.

Sharing this sensitive information between several users without compromising it, represents a challenge and an opportunity for innovators and staff who are concerned with the proper management of information that is growing exponentially. In the healthcare system, it is important to comply with the specific requirements that are within the Official Mexican Standard NOM-024-SSA3-2012 which mention that the Health Service Providers who use the Electronic Registration Information Systems (SIREs) must guarantee the confidentiality of the patients' identity as well as the integrity and reliability of the clinical information and establish the pertinent and adequate security measures in order to avoid the illicit or illegitimate use that may damage the legal sphere of the holder of the information [4].

This article presents a proposal for a solution through the application of the Shamir Threshold scheme, as a management system tool that allows sharing information-sensitive compliance with the guidelines of the changes in the Official Mexican Standard NOM-024-SSA3- 2012 using current trendy technologies such as NodeJS and Neo4j.

## **2 Secret Sharing Schemes**

In the scenario where several users are going to have access to the information that is encrypted to maintain their confidentiality and integrity to external attack, the problem of sharing the key for decrypting information without users having to intervene in process in a conscious manner is presented.

A possible solution to this problem is to use schemas such as division of secrets or the Shamir Threshold Scheme [5], in which a secret (in this case the decryption key) is split into individual secrets for each user system who wants access to information. In the scheme of division of secrets, it is proposed that a message  $M$  can be divided among  $m$  persons. To reconstruct the secret, it is necessary to gather all the pieces.

The Shamir Threshold scheme states that a message  $M$  can be divided among  $k$  person, but unlike the previous scheme it is possible to define a threshold of  $n$  required submessages or pieces to be able to reconstruct the original message without having to gather all the submessages. The minimum pieces are restricted to  $k \leq n$  [6].

In the other hand, firstly the owner of the information must authorize to the specific users into the collaborating network who will have access to the key. Then each time an owner generates new information, an encryption key will be provided to the users that are in the authorization list of the owner.

### **3 Related Work**

Nowadays exist different systems for share files in a user community like Dropbox® that protects the files in storage and in transit between applications and servers. The files data of Dropbox® in storage are encrypted with an encryption of 256 bits through the advanced encryption standard (AES). Dropbox applies secure socket layer protocol (SSL), transport layer security (TLS) for data transfer, what creates a more secure tunnel protected by the advanced encryption standard (AES) of 128 bits or higher. The generation, the exchange and the storage encryption keys are distributed to allow decentralized processing [7]. Besides, Google Drive protects users against any modification, divulgation, or unauthorized destruction of data by an unauthorized access. Encrypted using SSL protocol, offers the possibility to configure two steps verification to access Google accounts [8]. OneDrive files aren't shared with other people unless they are saved in a public folder or choose to share with specific person. This system saves multiple copies of each file in servers and different units, creates a secure password, and adds safety information to the account of Microsoft, also an additional security code is required each time you log on a new device or one that it is uses temporally [9]. All of them, depends mainly on the email provided and the password of this account. In all these cases, the owner shares the files or folder and once he shares it, the other user could see everything that is uploaded in this folder. MEGA uses point to point encryption unlike most cloud storage providers, and the owner controls who has access to his data, not even MEGA can access them [10]. As it is observed, these reservoirs protect the transmission and the files in cloud systems with the same level of security, not considered double check for specific files.

The Secret Sharing Scheme has been used by X.Huijuan, S. Wei and H.C. Hao in [11] to ensure security in cloud computing inside a cloud service provider, and this application shows that this technique is useful enough to protect all the sensitive data that will be stored in our proposed application. In [13], Peeters et al. proposed that the user's secret key could be distributed among a group of personal devices for authentication, demonstrating that this scheme has many other potential applications, not only secure data sharing, also demonstrated by Kaul et al [14]. On the other hand, Gan et al. uses this scheme with the generation of Lagrange polynomials interpolating to prevent the system from loss, damage, and external attack, reducing the key holder's responsibility [12]. It seems to be a more robust application comparing to the proposed in this paper. Nevertheless, the optimal combination of specific techniques for taking care of the information and stablish several levels of security is still an open research area.

## 4 Application of the Shamir Threshold Scheme in a Close Community

The proposed solution is described in this section. The sensitive information contained in the experimental clinical studies is encrypted using the Advanced Encryption Standard (AES) using a secret key that will be divided into  $m$  subkeys using the Shamir Threshold scheme, so the system administrator is able to assign a subkey to each user. They have the chance of recovering the original key so they can encrypt and decrypt information in the place and time they required it, without compromising it. The server verifies the key and gives access to the information, guaranteeing its security and confidentiality, thus complying with the provisions in Official Mexican Standard NOM-024-SSA3-2012.

This procedure also has the advantage that the users do not have to maintain a physical contact to agree the access key. A minimum threshold of 2 subkeys is proposed to get the secret key and thus the user could retrieve the encrypted information.

The developed system allows different users to share files among them in a secure way. The AES encryption standard was used, which is responsible for encrypting and decrypting such information and the Shamir Threshold scheme to divide the encryption key and assign a subkey to each registered user, including one more key to the server. Each user will have a pair of two partial key, the first one is for sharing the files with everyone in the community in a secure way and another key for access to the sensitive information under his responsibility, both must be completed with the Server partial key.

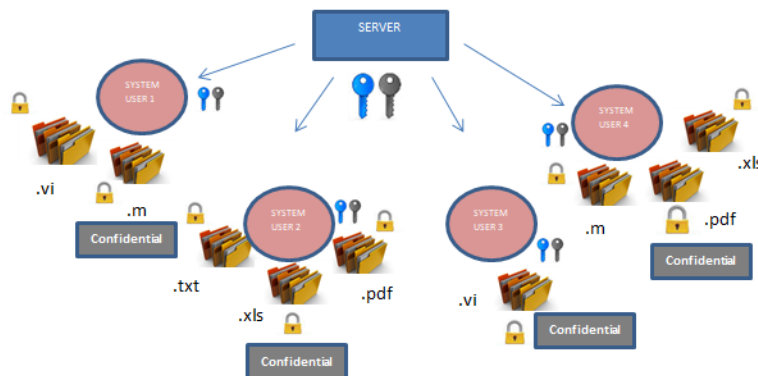


Fig. 1. General scheme of the community.

The operation of the system is divided into three main processes: a) User registry and validation process b) Configure the system for the first time and c) Data encryption and decryption.

### 4.1 User and Validation Process

In the moment, the user wants to be register into the system, he needs to be validated by the administrator and it be assigned a subkey in order to access to the encrypted

information in the collaborative community. Once it is recognized and validated, the subkeys of all the users stored in the database are obtained and used to generate a new one for this new user using the Shamir Threshold scheme (Fig 2).

The process follows these steps:

1. *Enter user data*: The user enters its data in the register form, then this information is sent to the server to be encrypted and stored.
2. *Generate user's private key*: A new random private key is generated to encrypt the user's sensitive data, thus, only the owner of the data has access to the information associated to this key.
3. *Save encrypted data*: The user's data is encrypted using AES-256 and the private key generated before and then it is stored in the database, with a null subkey field.
4. *User validation*: The administrator validates a public user's key for the user n.
5. *Request to generate subkey*: A new subkey is requested to the server.
6. *Generate subkey*: The requested subkey is generated using Shamir Threshold Scheme.
7. *Assign subkey*: The new generated subkey is assigned to the validated user, and the subkey is stored inside the user's database registry.

#### 4.2 Initialization of the Assignment of Subkeys.

This process is carried out just when the first user is registered (Fig. 3).

1. *Generate secret key*: The secret key that will be used in the encryption process is generated.
2. *Divide secret key*: The secret key generated before being split into two subkeys using Shamir Threshold Scheme.
3. *Store server subkey*: One of the generated subkeys is assigned to the server, thus, it is stored in it.
4. *Register the first user of the system*: The data of the first system user is stored in the system.
5. *Store first user subkey*: The left subkey is assigned to the first user and is stored in the user's database registry.

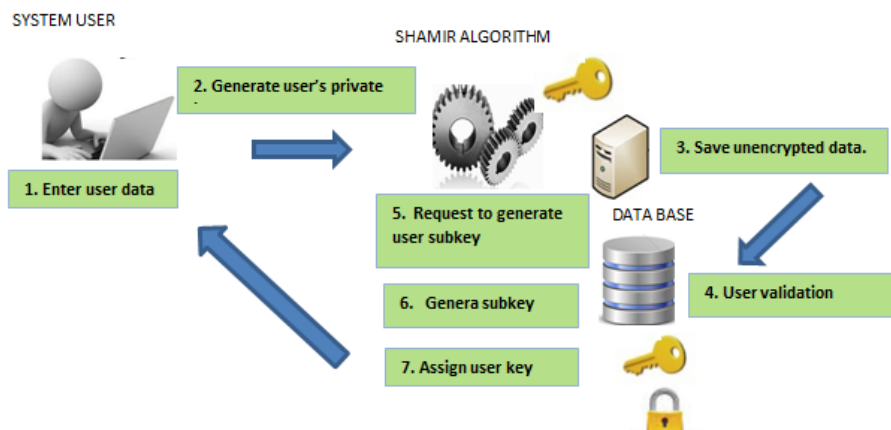


Fig. 2. User registry and subkey assignments.

### 4.3 Data Encryption and Decryption

This process is triggered when a user uploads new data file to the community. In Fig. 4 it is summarized this process.

1. *Upload file*: The user uploads the file to the server.
2. *Retrieve user identification and user subkey*: The user's and server's subkeys are retrieved to recover the secret encryption key and authorized the access to the information.
3. *Recover secret encryption key*: The secret encryption key is recovered by joining the two retrieved subkeys with the use of the Shamir Threshold Scheme.
4. *Encrypt file*: The file is encrypted with AES using the secret key recovered.
5. *Store encrypted File*: The encrypted file is stored in the system, and its URL, the relation with the user and the metadata are stored in the database.

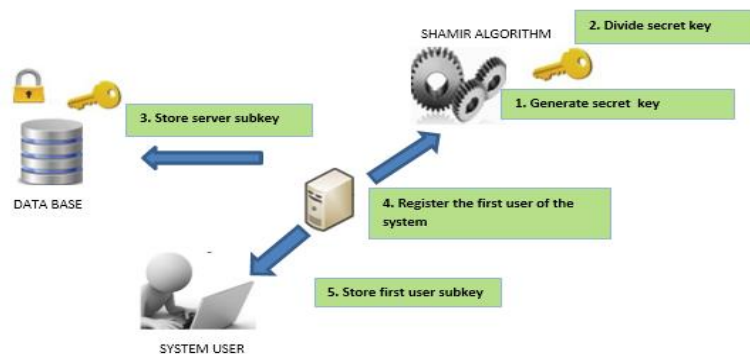


Fig. 3. Initialization of the system.

Once the user had already stored information, he must follow the procedure shown in Fig. 5 to recover the information. If a community member requires to download the information, the first step is to locate the requested resource, recover the encryption key and then the decrypting process is performed and finally the file is downloaded. If he intended to open a sensitive information of another member, this key will not be useful and the decrypting process is not performed. If the community member is the responsible of the confidentially information, his key will be recognized and he will able to recover the decrypted information.

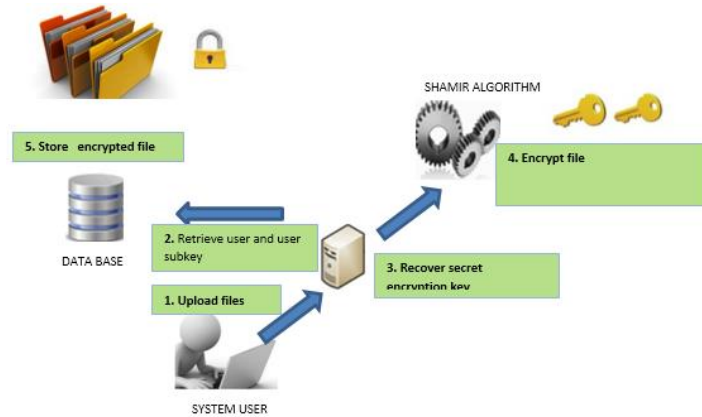


Fig. 4. Information upload and encrypting.

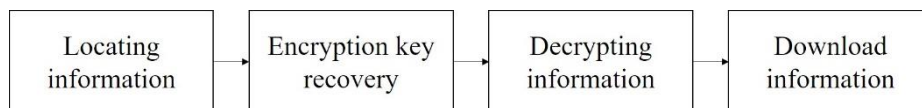


Fig. 5. Information download and decrypting.

## 5 Preliminary Results

The system was tested in a personal computer running Windows 10 with 8GB RAM and an AMD 4Core A10-8700P, up to 3.2 GHz processor. A randomly secret key was generated to provide the subkeys to the users using the Shamir Threshold scheme.

Besides, the AES symmetric encryption standard for file encryption and some current technologies such as Node.js and Neo4j were used for generating a system with the architecture shown in Fig. 5.

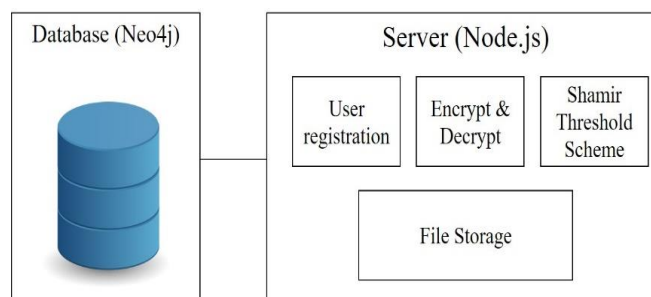


Fig. 6. System architecture.

As shown in Fig. 6, the system is composed by two main modules, the server and the database. The server is responsible for encrypting and decrypting sensitive information and files, running the Shamir Threshold Scheme processes and storing the encrypted files on the database.

In the database, the sensitive information of each registered user is saved including the corresponding private key, the file's meta data, and the relationships that link the files with their respective owner. The relationship of the owner with its files are exemplified in Fig. 7.

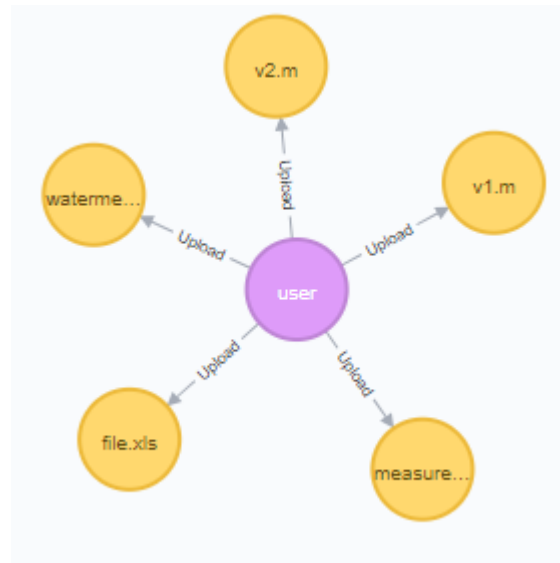


Fig. 7. Example of one Generated User Node

The server uses the Shamir Threshold scheme to split the primary secret key and generate a subkey for each user, following the procedure described in Fig. 2 and Fig. 3.

	Key	Value	
<input checked="" type="checkbox"/>	username	testuser	<pre> {   "storedUser": {     "id": 94,     "name": "Testuser",     "surname": "Testuser",     "email": "testuser@email.com",     "username": "testuser",     "role": "ROLE_ADMIN",     "password": "\$2a\$10\$1bDpJ9L6nM9Lc",     "subkey": null   } } </pre>
<input checked="" type="checkbox"/>	name	TestUser	
<input checked="" type="checkbox"/>	surname	TestUser	
<input checked="" type="checkbox"/>	password	testUser	
<input checked="" type="checkbox"/>	email	testUser@email.com	

Fig. 8. Left: User registration section of the database; Right: Server answer after being validated the new user by the administrator.

In the left of Fig. 8 it is shown an example when the user data is sent to the server for registration, and in the right of Fig. 8 is shown how the server response. An object

containing the data of the user registered in the system database without validation. The user does not yet have a subkey which it is observed because the object has a null value in the password record until his data are verified. Once a user is validated, the null value that contains the subkey field is overridden with a subkey obtained from the Shamir Threshold subsystem as shown in Fig. 9.

With this subkey users can retrieve the original key, that will be used to encrypt the files that he has uploaded to the system and then other users will be able to decrypt files if they have permission to retrieve the original encryption key (Fig. 10).

```
{
  "validatedUser": {
    "id": 94,
    "name": "Testuser",
    "surname": "Testuser",
    "email": "testuser@email.com",
    "username": "testuser",
    "role": "ROLE_ADMIN",
    "password": "$2a$10$1bDpJ9L6nM9Ld8Pr5",
    "subkey": "8071fd22b39c99420b55cc4d9645"
  }
}
```

Fig. 9. Validated new user by the administrator.

In the Fig. 10, in the right image is presented the result for a validated member trying to download and read his own sensitive information file. In the left image is presented the result of a not validated member, trying to read a confidential file from other user.

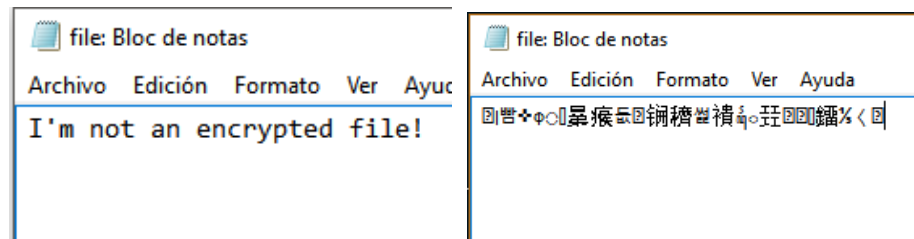


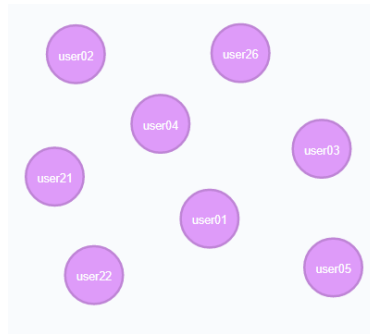
Fig. 10. Download results with, a validated member of the community (right image) and non-authorized member (left image).

The results shown in Fig. 10 prove that the encryption of a file becomes unreadable and impossible to interpret correctly it. In this system, encryption ensures its content, since the user cannot retrieve it if he does not have the encryption key, so, it is possible to expect that the presented system allows to create a community of users that share information securely with each other, sharing the same encryption key that it is divided in the number of users belonging to that community. And with another private key for his confidential files.

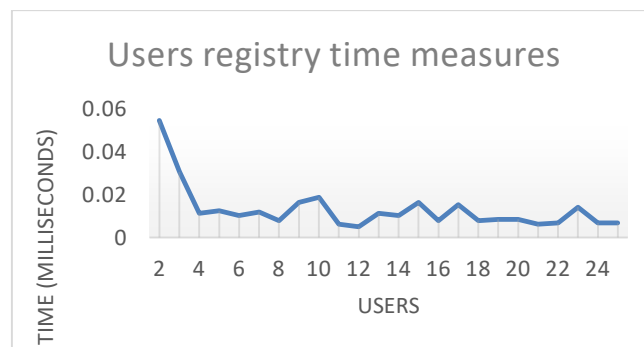
As a validation proof of the Shamir scheme used as a security proposal, we used a community of 25 registered users as shown in Fig. 11, each user uploaded 8 files. We

test systematically that only users belonging to the community with an appropriate sub-key, can access the encrypted information and in 100% of cases it was achieved.

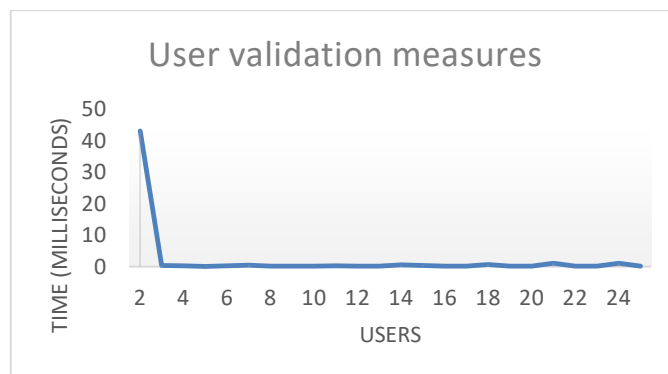
Also, the times taken by registering and validating a new user were measured in order to know the relationship between these parameters. As shown in Fig. 12, the users registry time seems to be lower at 25 users than at only 1 or 4 users, but, all these measures depend on the resources if the used server. The average of time was of 0.01787384 milliseconds.



**Fig. 11.** Part of the designed community in the database.



**Fig. 12.** Users registry time measures.



**Fig. 13.** Users validation time measures.

Fig. 13 shows a similar situation than Fig. 12. Although the measures depend on the server's hardware resources use, it is observed a stabilization of this parameter. Having an average time of 0.414505675 milliseconds. This measurement was computed after the initialization procedure and validation of the second new user because a different procedure is followed in the initialization of the system so that's why the measures of only one user does not appear in the graphics.

It can be concluded that the system can handle and generate 25 and more subkeys without any problem.

## 5 Conclusions

The use of the Shamir Threshold Scheme in conjunction with AES let a collaborating community of user safely share information among them, but with a private key for confidential files, let the availability of this sensitive information just to the designated responsible, complying with the provisions of Official Mexican Standard NOM-024-SSA3- 2012. Promising results are obtained of the implementation of this type of techniques to achieve security for different kind of information into the proposed system because helps to handle data files in a laboratory with interdisciplinary projects that use the Internet to keep in touch among the users of the system. It is still necessary to carry out future work in this application, such as defining the methodology for sharing information between communities, and inserting it into the complete information management system, in order to be able to demonstrate its concrete effectiveness.

## References

1. Hey, T., Trefethen E.A.: The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems* 18(8), 1017–1031 (2002)
2. Arbona, A., Benkner, S., Engelbrecht, G., Fingberg, J., Hoffman, M., Kumpf, K., Guy, L., Woehrer, A.: A Service-Oriented Grid Infrastructure for Biomedical Data and Compute Services. *IEEE Transactions on NanoBioscience* 6(2), 136–141 (2007)
3. Secretaría de Salud: NORMA Oficial Mexicana NOM-024-SSA3-2012, Sistemas de información de registro electrónico para la salud. Intercambio de información en salud, <http://www.dgis.salud.gob.mx/descargas/pdf/NOM-024-SSA3-2012.pdf> (2012)
4. Malin, B.A., El, E.K., O'Keefe, C.M.: Biomedical data privacy: problems, perspectives, and recent advances. *Journal of the American Medical Informatics Association* 20(1), 2–6 (2013)
5. Washington, L.C.: *Introduction to cryptography: with coding theory*. New Jersey, United States: Pearson Education (2006)
6. Shamir, A.: How to share a secret. *Commun. ACM* 22(11), 612–613 (1979)
7. Dropbox, Inc.: Security - Dropbox, <https://www.dropbox.com/security#files> (2017)
8. Google, Inc.: Privacy Policy-Privacy and Terms. <https://www.google.com/intl/policies/policies/privacy/> (2017)
9. Microsoft: Seguridad de archivos de OneDrive, <https://support.office.com/es-es/article/Seguridad-de-archivos-de-OneDrive-23c6ea94-3608-48d7-8bf0-80e142edd1e1?ui=es-ES&rs=es-HN&ad=US>. (2017)

10. Mega Limited: MEGA, <https://mega.nz/>
11. Huijuan, X., Wei, S., Hao, H.C.: Shamir's threshold scheme to Ensure Security in Cloud Computing. *Applied Mechanics and Materials*, vol. 543, pp. 3632–3635 (2014)
12. Gan, X., Liu, B.: Shamir Threshold Based Encryption. *Applied Mechanics and Materials*, vol. 52, pp. 709–712 (2011)
13. Peeters, R., Singelée, D., Preneel, B.: Towards More Secure and Reliable Access. *IEEE Pervasive Computing* 11(3), 76–83 (2012)
14. Kaul, S.D., Awasthi, A.K.: Privacy Model for Threshold RFID System Based. *Wireless Personal Communications*, vol. 95, pp. 2803–2828 (2017)

# Proposal of a Communication Architecture for the Configuration and Monitoring of an Electric Microgrid

V. H. Garcia<sup>1</sup>, R. Ortega<sup>1</sup>, R. Hernández<sup>2</sup>, M.A. Ramírez<sup>1</sup>

<sup>1</sup> Instituto Politécnico Nacional, ESCOM, Mexico City, Mexico

<sup>2</sup> Instituto Politécnico Nacional, UPIITA, Mexico City, Mexico  
 vgarciaortega@yahoo.com.mx, rortegag@ipn.mx,  
 rhtovar@ipn.mx, ancerritos22@gmail.com

**Abstract.** This paper proposes the use and application of a wireless technology to create a communication architecture for the configuration and monitoring of an electric microgrid. The architecture uses a wireless sensors network with sensor nodes based on a digital signal controller, DSPIC30F4013, which have a unit to execute algorithms for digital signal processing and a 16-bit microcontroller with several peripherals. The sensor nodes have standard PMOD and MikroBus connectors for sensor interface and communication modules. The sensor nodes use a WIFI communication module which has a 32-bit embedded processor where the TCP/IP protocol stack is located. In the communication architecture, an embedded system based on an ARM Cortex A53 quad-core processor is used as a server. The server uses a client-server architecture using TCP sockets and a Linux custom kernel version created using Yocto Project.

**Keywords:** microgrid, Embedded system, Sensor network, Digital signal controller.

## 1 Introduction

During recent years, a great interest in renewable energy sources has presented. These sources do not disturb the environment like conventional, fossil fuel-based energy sources do. Several kinds of renewable energy sources exist, among which wind power and solar energy stand out. Nowadays, the production via renewable energy sources is a feasible option and its proposed to be used along with the existing generation and distribution schemes. By doing this, it is sought to promote and diversify the energy supply so that, in the future, they may play an important role in the new technological-environmental electric energy generation schemes [1]-[3].

An essential point to the use of renewable energy sources is the need to implement interfaces which allow these sources to connect to the electric network, as well as to feed electric loads directly. Such interfaces are known as microgrids [4]. Another important point to consider in this new generation scheme is the flexibility and autonomy these microgrids work with. Namely, in case of failures on the distribution net, they can provide power directly to the user, achieving a higher flexibility than current distribution schemes. This new generation scheme is known as Distributed Generation (DG) [5]. Wired and wireless communication technologies can be used in microgrids [6]. The most popular wired technologies used in power systems are the RS-

232/422/485 serial communication nets with bus technologies (like Modbus, Profibus and CANbus) [7], the power line communication [8] and Ethernet [7], [9]. On the other hand, the most popular wireless technologies used in power systems are the Global System for Mobile Communication (GSM) [7], WIFI [11], [12], WiMax [8], ZigBee [6], [13]-[15], Bluetooth[14], radiofrequency [14] and Microwave [16].

Wired technologies have large bandwidth for data transfer and are more dependable than wireless technologies, however the installation cost is relatively high. Whereas wireless technologies have low installation cost and are better suited to remote areas. Furthermore, they are better suited to future expansions.

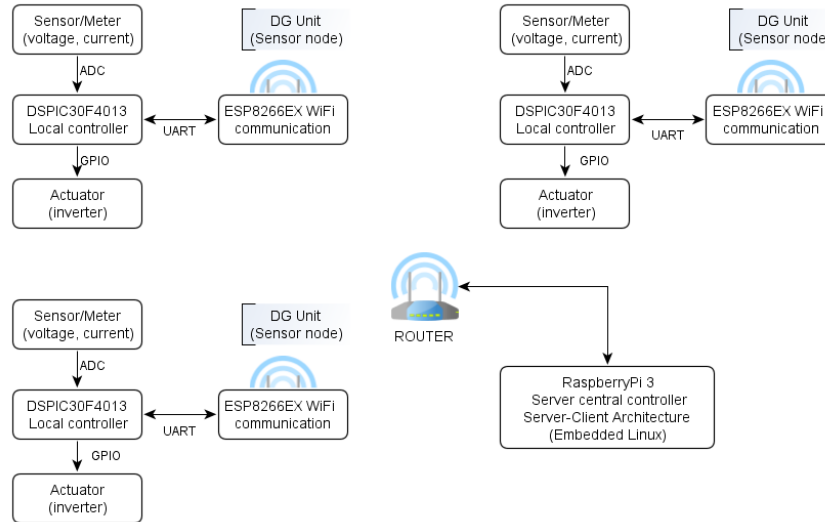
A comparison between different wireless transmission technologies is done in [6] showing that WIFI uses the highest data transfer rate at 54Mbps, however the initial cost is modest compared to other technologies. Nowadays, this cost has considerably decreased in such a way that is considered to be low. The ESP8266EX [17] is a System On Chip with efficient energy consumption handling, compact design and a dependable performance on Internet of Things (IoT) applications. It has a WIFI controller, with a 32-bit processor where the TCP/IP protocol stack is located, embedded on the same integrated circuit. Also, it has SPI, UART and I2C communication interfaces to operate with any given microcontroller. This SoC is used on Wireless Sensor Networks for control and monitoring. A WSN is a wireless network made up of autonomous sensors that are implemented in areas of interest and have common features such as: data processing, storage capacity, wireless communication interfaces and limited power consumption. These networks are used to control and monitor several application types on different environments [18].

In this work, the use of wireless technology is proposed to create a communication architecture that uses a WSN based on a DSPIC30F4013 [19] Digital Signal Controller (DSC) to process the control algorithms used during the operation of monophasic inverters on island mode [20] and on a ESP8266EX SoC to communicate with a WIFI network. The SoC configuration is done through AT commands using the UART interface of the DSC. The data is sent to a central controller that is located on an embedded system based on a quad-core ARM Cortex A53 processor, RaspberryPi 3. The server which receives the data from the WSN is located on this embedded system.

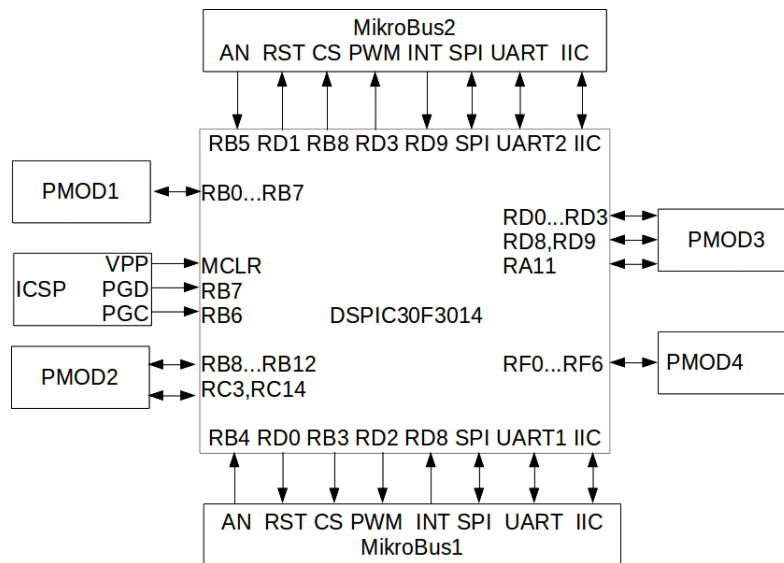
## 1.1 Microgrid Communication Architecture

The microgrid communication architecture is shown in Fig. 1. This architecture is constituted by three modules:

- Sensor Node
- Communication Module
- Server Module



**Fig. 1.** ¡Error! Secuencia no especificada. Proposal of a Communication Architecture for a Microgrid.



**Fig. 2.** Sensor Node.

## 1.2 Sensor Node

This sensor node (SN) is based on the DSPIC30F4013. Among its peripherals, this DSC has [19]: 2 UART, 1 SPI, 1 I2C, 5 TIMERS, 4 PWM, 1 CAN, 1 DCI, 48Kbytes of program memory and 2Kbytes of data memory. The DSC has a 16-bit RISC

architecture. The SN uses two connectors based on the MikroBus standard [21] and four connectors based on the PMOD interface specification [22] to smoothly adapt peripheral modules. A 14745600 Hz crystal is used on the SN, with which 29.4912 MIPS are reached. Additionally, the In-Circuit Serial Programming (ICSP) connector and its 5V and 3.3V power source are owned. The MikroBus standard essentially converts the SPI, UART and I2C communication interfaces while the PMOD handles every input-output terminal.

The SN node performs the processing of the control algorithms used in the operation of monophasic inverters on island mode as part of a microgrid. The monitoring of different variables and parameters is also performed for this mode of operation such as: voltage, current and harmonic measurement. The voltage is 127V and the current is 3.4A for a 400W inverter [20]. The SN interconnection architecture is shown in Fig. 2.

### 1.3 Communication Module

The communication module is based on a ESP8266EX SoC that uses a UART interface to communicate with the DSC. The SoC has a 9,600 bauds rate configured by default with a minimal frame composed of one start bit, one stop bit and 8 bits per data with no parity bit used. That is 10 bits per frame. The transfer rate is given by (1):

$$bytes_{transfer} = \frac{BaudRate}{BitsPerFrame}. \quad (1)$$

Given a 9,600 bauds rate, the transfer rate is 960 bytes per second. The sampling frequency used in the control of the operation of the monophasic invertors on island mode is 40 KHz with a resolution of 16 bits per data. Given this sampling frequency and a word size of 16 bits, a transfer rate of 40,000 words per second is needed to send samples of the input variables in real time. With a rate of 921600 bauds, a transfer rate of 92160 bps is attained, namely, 46,080 words per second. Using this speed, the samples can be sent in real time.

To configure the UART of the DSC to a speed of 921,600 bauds, we use equation (2) in order to find the right value to use in the UxBRG register:

$$UxBRG = \frac{FCY}{16 * BaudRate} - 1. \quad (2)$$

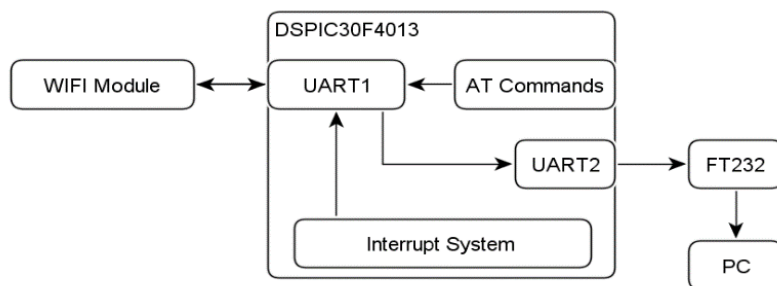
With the 14,745,600 Hz crystal located in the SN, a 8-bit PLL is configured so that FCY is 29.4912 MHz. Given this value for FCY, we get UxBRG = 1. Using this crystal value, the desired baud rate is achieved. The DSC communicates with the Wi-Fi SoC via AT commands using the UART1 interface with an initial rate of 9,600 bauds. To receive the command response, the UART1 receiving interruption is enabled. The commands used on this application are shown in Table 1.

To debug the command execution between the DSC and the Wi-Fi SoC, the UART2 interface is used to send every response from each command through a FT232 serial-USB converter. The UART2 is configured at a rate of 9,600 bauds. On the computer, the response acquisition is done by programming the UART through the File Hierarchy System (FHS) using a program written in C language. The architecture of the proposed test application is shown in Fig. 3.

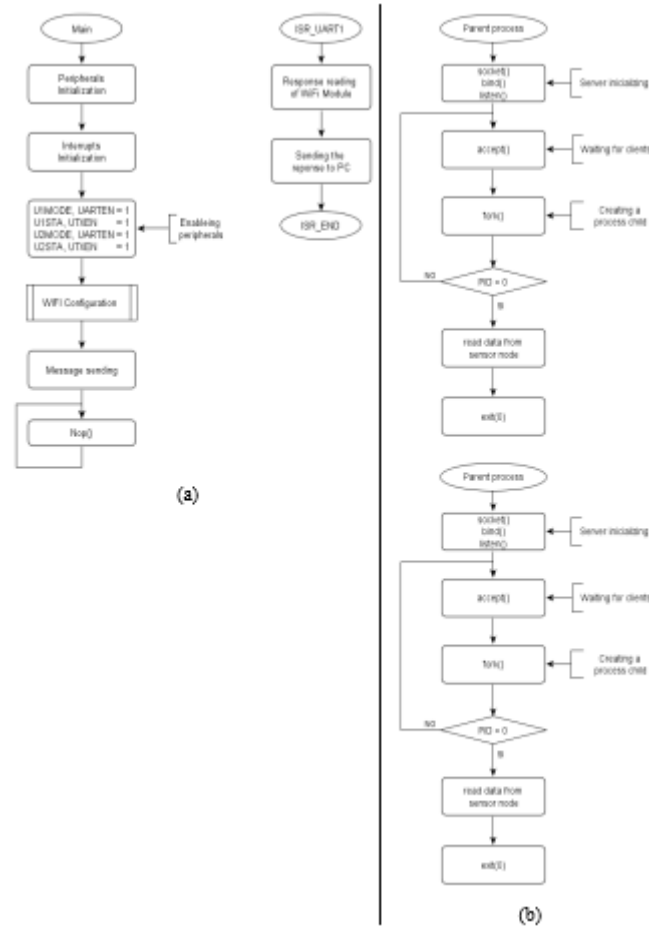
**Table 1.** AT Commands for configuring ESP8266EX.

Function	AT Commands	Response
Restart	AT+RST	OK [System Ready, vendor:www.ai-thinker.com]
WiFi Mode	AT+CWMODE? AT+CWMODE=1 AT+CWMODE=2 AT+CWMODE=3	Query STA Mode AP Mode Both
TCP/UDP Conections	AT+ CIPMUX? AT+ CIPMUX=0 AT+ CIPMUX=1	Query Single Multiple
JoinAccess Point	AT+CWJAP? AT+CWJAP= "SSID","Password"	Query AT+CWJAP?+CWJAP: "Router" OK
Get IP Address	AT+CIFSR	AT+CIFSR 192.168.0.105, OK
Set up TCP/UDP conection	AT+CIPSTART=? (CIPMUX=0) AT+CIPSTART = <type>,<addr>,<port> (CIPMUX=1) AT+CIPSTART = <id><type>,<addr>,<port>	Query id = 0-4, type = TCP/UDP, addr = IP address, port= port
Set Buffer Size	AT+CIPSEND=N	N, Number of bytes to send
Baud Rate	AT+UART=BAUDRATE	

The whole application is developed in a C language interface for embedded systems and the assembly language of the DSC. The flowchart of the proposed system for the Wi-Fi module in shown in Fig. 4 (a).



**Fig. 3.** Testing Unit for the Sensor Node.



**Fig. 4.** (a) Flowchart for WiFi Module testing, (b) Flowchart for Server Application.

#### 1.4 Server Module

The server module is implemented on the development board RaspberryPi 3, whose main characteristics are [24]:

- 1.2 GHz quad-core Cortex A53 processor with dedicated 512Kbyte L2 cache in BCM2837 SoC.
- 1GB RAM
- VideoCore IV 3D graphics core.

The operating system used with the development system is a personalized version, called rpi-basic-image, generated in Yocto Project [25]. This version has the hardware configuration needed to initialize the Linux kernel on the RaspberryPi with some additional features using only 180Mb. The kernel version used is 4.4.32. The boot system used for the server application daemon is System D. The server application is

programmed in C language and uses Berkeley TCP sockets in a client-server architecture. The application uses the fork() system call to create a process for every client that requests a connection to the server, allowing process-level parallelism. The flowchart for the server application is shown in Fig. 4 (b).

## 2 Testing and Results

A sensor node was designed, fabricated and assembled with a DSPIC30F4013 DSC to implement local control and to configure the ESP8266EX Wi-Fi SoC. A Wi-Fi 3 Click module from Mikroelektronika vendor was used. Fig. 5 shows the sensor node working with the SoC. This figure also shows the responses to the AT configuration commands.



Fig. 5. Sensor node in operation with ESP8266EX.

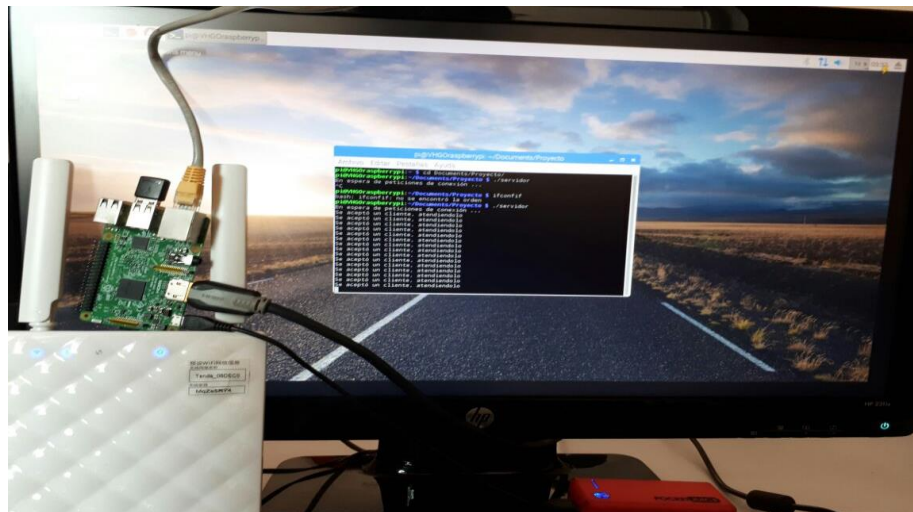
To verify the proper operation of the server on the embedded RaspberryPi 3, system tests were carried out by connecting from 1 to 30 clients on the server simultaneously. Each client sent a buffer of 100,000 64.bit integer data at different frequencies, varying from 1KHz to 40KHz, with increments of 5KHz. The obtained results are shown in Table 2.

In Table 2, the first column shows the frequency at which each client sent the data. The second column shows the number of clients connected to the server. The third and fourth columns show the time the server took to receive all the client data and the last column shows the number of buffers received from the clients on the server.

The obtained results show that from 15 simultaneously connected clients, disconnections may occur and data buffered may get lost. Fig. 6 shows the execution of the server.

**Table 2.** Results of system tests of the server application.

Frequency (khz)	Clients	t (m)	t (s)	BUFFERS	Frequency (khz)	Clients	t (m)	t (s)	BUFFERS	Frequency (khz)	Clients	t (m)	t (s)	BUFFERS
1	1	1	59.7	1	15	1	0	8	1	30	1	0	5.411	1
	5	2	45.6	5		5	0	8.77	5		5	0	4.75	5
	10	3	41.04	10		10	0	10.66	10		10	0	5.39	10
	15	4	52.48	13		15	0	14.43	13		15	0	8.138	15
	20	8	15	20		20	0	21.549	17		20	0	8.35	17
	25	8	37.2	21		25	0	26.21	21		25	0	12.72	21
5	30	10	13.59	25	20	30	0	35.614	26	35	30	0	18.35	27
	1	0	30.94	1		1	0	6.13	1		1	0	4.17	1
	5	0	26.37	5		5	0	6.68	5		5	0	4.59	5
	10	0	35.89	10		10	0	9	10		10	0	5.8	10
	15	0	42.31	12		15	0	12.65	15		15	0	6.33	13
	20	1	5.48	20		20	0	14.7	17		20	0	9.22	19
10	25	1	26.4	22	25	25	0	19.25	21	40	25	0	10.66	20
	30	1	45.31	25		30	0	20.5	23		30	0	16.32	29
	1	0	12.29	1		1	0	4.84	1		1	0	3.7	1
	5	0	13.44	5		5	0	5.76	5		5	0	3.5	5
	10	0	14.48	8		10	0	6.65	9		10	0	3.96	10
	15	0	24.4	14		15	0	8.92	12		15	0	4.41	12
	20	0	34.9	18		20	0	11.35	17		20	0	8.54	20
	25	0	46.25	22		25	0	16.38	24		25	0	11.66	22
	30	1	51	24		30	0	19.34	27		30	0	10	27



**Fig. 6.** Server in operation.

### 3 Conclusions and Future Work

The sensor node designed with the DSPIC30F4013 can process up to 29.4912 million instructions per second with the 14.7456 oscillator selected for the SN with a PLL of 8, this data is provided by the manufacturer. This oscillator value was selected in order to reach a 921600 baud rate on the UART with 0% configuration error. Given this transfer rate for the UART, transfer rate of over 40KHz can be reached. This node can process the control algorithms used on the GD inverters since it contains a DSP on the DSC. Furthermore, on the DSC the module ESP8266EX is configured to gain connectivity to the wireless sensor network with a data rate of 54 Mbps. This data rate is higher than the data rate presented with GSM [7], WiMax [8], ZigBee [6], [13] – [15], Bluetooth

[14] and radiofrequency [14]. Furthermore, the module ESP8266EX presents a low cost nowadays.

The microgrid server is located on a quad-core ARM Cortex A53 based embedded system and a personalized linux version created using Yocto project for improve the memory space. This linux version occupies only 180Mb and allows only for the required processes for the server application to be running.

Using the DSC and the server embedded system, a low cost, robust communication architecture is obtained. This architecture uses only two communication layers. The first layer contains the sensor nodes and the second layer contains the embedded server. This architecture avoids the need to use multiple communication layers in microgrids as described in [6], [13] – [15].

This communication architecture can be used for processing control algorithms used in the operation of monophasic inverters on island mode as part of a microgrid.

Furthermore, this architecture allows to measure the phase of different inverters with the purpose to connect them in parallel and feed a greater load.

**Acknowledgements.** The authors would like to thank the Postgraduate and Research Division of the National Polytechnic Institute who contributed to the development of this work through the SIP20170127 multi-disciplinary project.

## References

1. Mikati, M., Santos, M., Armenta, C.: Modelado y Simulación de un Sistema Conjunto de Energía Solar y Eólica para Analizar su Dependencia de la Red Eléctrica. *Revista Iberoamericana de Automática e Informática Industrial*, Vol. 09, pp. 267–281 (2012)
2. Pastora, R., Leão, S., Luiz, F., Antunes, M., Greison, T., Lourenço, M., Kleber Rodrigues, K.: Uma Visão sobre a Integração à Rede Elétrica da Geração Eólio-Elétrica. *Revista IEEE América Latina* 7(6), 620–629 (2009)
3. Yingjun, R., Qingrong L., Weiguo, Z., Ryan, F., Weijun, G., Toshiyuki, W.: Optimal option of distributed generation technologies for various commercial buildings. *Applied Energy Journal* 86(9), 1641–1653 (2009)
4. Kyriakarakos, G., Dounis, A., Rozakis, S., Arvanitis, K., Papadakis, G.: Polygeneration microgrids: A viable solution in remote areas for supplying power, potable water and hydrogen as transportation fuel. *Applied Energy Journal* 88(12), 4517–4526 (2011)
5. Manfren, M., Caputo, P., Costa, G.: Paradigm shift in urban energy systems through distributed generation: Methods and models Review. *Applied Energy Journal* 8(4), 1032–1048 (2011)
6. Setiawan, M.A., Rajakaruna, S.: ZigBee-Based Communication System for Data Transfer Within Future Microgrids. *IEEE Transactions on Smart Grid* 6(5) (2015)
7. Radhika, N., Vanitha, V.: Smart grid test bed based on GSM. *Procedia Engineering Journal* 30(211), 258–265 (2012)
8. Yoon, S., Jang, S., Bahk, S., Kim, Y.: Opportunistic routing for smart grid with power line communication access networks. *IEEE Transactions on Smart Grid* 5(1), 303–311 (2014)
9. Zhang, C., Ma, W., Sun, C.: A switchable high-speed fiber-optic ring net topology and its method of high-performance synchronization for large-capacity power electronics system. *International Journal of Electric Power & Energy Systems*, vol. 57, pp. 335–349 (2014)
10. Zhang, Y. *et al.*: Distributed intrusion detection system in a multi-layer network architecture of smart grids. *IEEE Transactions on Smart Grid* 2(4), 796–808 (2011)

11. Usman, A., Shami, S.H.: Evolution of communication technologies for smart grid applications. *Renewable and Sustainable Energy Reviews Journal*, vol. 19, pp. 191–199 (2013)
12. Gómez-Cuba, F., Asorey-Cacheda, R., González-Castaño, F.J.: Smartgrid last-mile communications model and its application to the study of leased broadband. *IEEE Transactions on Smart Grid* 4(1), 5–12 (2013)
13. Liu, T. *et al.*: A dynamic secret-based encryption scheme for smart grid wireless communication. *IEEE Transactions on Smart Grid* 5(3), 1175–1182 (2014)
14. Langhammer, N., Kays, R.: Performance evaluation of wireless home automation networks in indoor scenarios. *IEEE Transactions on Smart Grid* 3(4), 2252–2261 (2012)
15. Batista, N.C., Melício, R., Matias, J.C.O., Catalão, J.P.S.: Photovoltaic and wind energy systems monitoring and building/home energy management using ZigBee devices within a smart grid. *Energy Journal*, vol. 49, pp. 306–315 (2013)
16. Deng, C. *et al.*: Terrestrial-satellite hybrid backbone communication network for smart power grid. *Energy Procedia Journal*, vol. 12, pp. 27–36 (2011)
17. Espressif Systems: ESP8266EX Data sheet, Version 5.3 (2016)
18. Fernandez-Berni, J., Carmona Galán, R.: *Vision-enabled WSN Nodes: State of the Art*. Elsevier (2012)
19. Microchip Technology Inc.: *DSPIC30F4013 Data Sheet, High Performance Digital Signal Controllers* (2004)
20. Ortega, R., Carranza, O., Sosa, J.C., García, V., Hernández, R.: Diseño de controladores para inversores monofásicos operando en modo isla dentro de una micro red. *Revista Iberoamericana de Automática e Informática industrial* vol. 13, pp. 115–126 (2016)
21. Mikroelektronika Inc.: *mikroBUS, Standard specifications* (2015)
22. Digilent Inc.: *Digilent PMOD Interface Specification* (2011)
23. Microchip Inc.: *ICSP Guide* (2003)
24. Raspberry Pi Foundation: *Raspberry Pi 3 Model B* (2017)
25. Yocto Project: *Yocto Reference Manual* (2017)

# Proposal of Architecture for the Monitoring of Vital Signs based on Embedded Systems

Jorge Martínez, Víctor García, Rubén Ortega

Instituto Politécnico Nacional, Escuela Superior de Cómputo,  
Sección de Estudios de Posgrado e Investigación, Mexico City, Mexico  
jmartineza0809@alumno.ipn.mx, vgarciao@ipn.mx,  
rortegag@ipn.mx

**Abstract.** In this paper, we propose an embedded system architecture, using programmable devices and standard sensors, to continuously monitor the vital signs of hospitalized patients. The vital signs to monitor are body temperature and heart rate. Body temperature is measured by the 16-bit MAX30205 infrared temperature IIC sensor. The heart rate is measured using an analog photoplethysmograph model SEN11574, which signal will be digitized using the 12-bit ADC MCP3201 with SPI interface. An embedded system based on the Xilinx MicroBlaze processor, which is a soft-core processor, implemented on a FPGA of Spartan 6 family is used. The processor local bus is used for the interface with the peripherals, an IIC core for the temperature sensor configuration and an SPI core for the ADC interface. Different algorithms for calculating the heart rate signal frequency are analyzed.

**Keywords:** embedded system, FPGA, sensors, signal processing, vital signs.

## 1 Introduction

In hospitals in Mexico, control of information on the health status of patients is of paramount importance for physicians and nurses to perform their care tasks efficiently. Usually, the control of information is carried out through a clinical file which, according to Official Mexican Standard NOM-168-SSA1-1998, is the "set of written, graphic and imaging documents or of any other kind, in which the health personnel, must make the records, annotations and certifications corresponding to their intervention, according to the sanitary dispositions" [1]. As part of the documents that make up this file, we find nursing sheets that, among other data, contain personal information of the patient and results of a physical examination or measurement of vital signs [2].

Nowadays, medical personnel must physically go to check the health status of each patient. The periodicity of the revision of the vital signs is given by the normativity of each institution and by the medical orders for each patient. Nursing records are made manually from a standardized printed format; they are stored in the patient's clinical record or, if the hospital handles an electronic medical record, the data in this record must be captured in the system.

The time required for the review of each patient may be prolonged because each of the vital signs must be verified in a separate way. The time required for the transcription

of data in nursing sheets should also be considered, as well as the possibility of making mistakes in filling the sheets of nursing.

## 1.1 Vital Signs

The four vital signs represent a simple assessment of the physiological and physical state of an individual. They are used by physicians as a quick and general evaluation of their patients and are often measured by a medical assistant prior to the medical office visit. The vital signs of hospitalized patients are also regularly measured for periodic evaluation of their condition. The four basic signs are: body temperature, heart rate, respiratory rate and blood pressure (see Table 1) [3].

**Table 1.** Definition and characteristics of vital signs.

Vital sign	Definition	Method of measurement	Stability range (people between 19 and 26 years)
Body temperature	Measurement of the kinetic energy of the atoms or molecules of a substance or object [4]	Contact thermometer (mercury or electronic)	Between 36.5°C and 37.2°C
Blood pressure	The force exerted by the blood against the walls of the arteries [5]	Tensiometer and a stethoscope	Systolic < 120 mm Hg Diastolic < 80 mm Hg
Heart rate	The number of times the heart beats per minute [6]	Pressing firmly on the arteries / Electrocardiogram	Between 60 and 100 beats per minute
Breathing frequency	The amount of breaths a person does per minute [7]	Count the breath for one minute	Between 12 and 16 breaths per minute

## 2 Proposed System Architecture

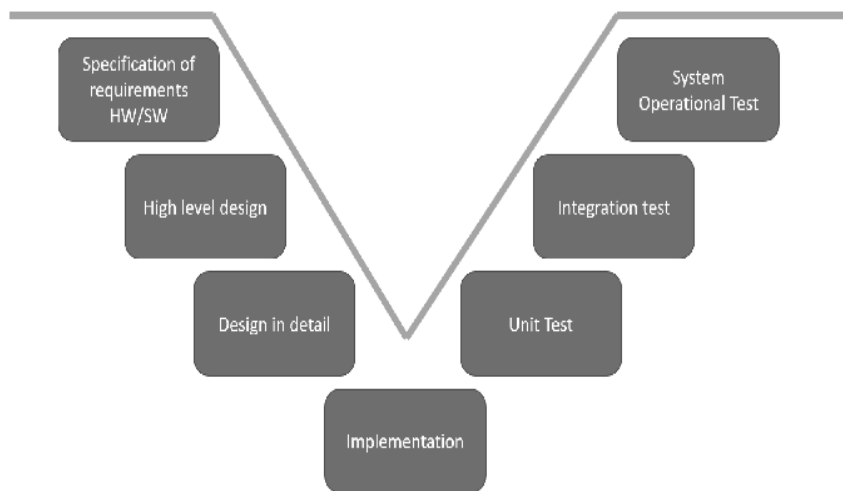
### 2.1 Methodology

In this research, a study from the general to the particular is realized, in other words, the deductive method is used [8]. It is from the investigation of monitoring systems and vital signs to the investigation of the characteristics and principle of operation of the sensors for their measurement, as well as the processing of the signals obtained.

The choice of vital signs to monitor is performed qualitatively while the choice of sensors and devices for their processing is performed quantitatively, considering the sensitivity, working range and processing speed of these devices, among other characteristics.

## 2.2 V-Model for Embedded Systems

The chosen model for the development of this embedded system is the V-model since the design stages can be related to the test stages. This is convenient for the type of systems involving hardware and software. For this, it is necessary to consider all the steps shown in Fig. 1 [9].



**Fig. 1.** V-model for the development of embedded systems.

## 2.3 System Analysis

The temperature sensor used is the MAX30205, which is powered with a voltage between 2.7 V and 3.3 V and with an average consumption of 600  $\mu$ A. The working range of this sensor is between 0 °C and 50 °C, with an accuracy of 0.1 °C. The output signal is digital with a 16-bit resolution. This sensor has an IIC interface, which is used for communication with the programmable device that processes the signal. The resolution of the output signal is sufficient to comply with the medical standard of the National Center for Technological Excellence in Health (CENETEC) which requires a minimum of 12-bit resolution [10].

The pulse sensor used to obtain the heart rate is the photoplethysmograph SEN-11574, due to the integration of a circuit containing a protection diode in the supply voltage, in addition to a filter and an amplifier to increase the amplitude of the pulse wave and normalize the signal around a reference point. The schematic diagram of this sensor is shown in Fig. 2.

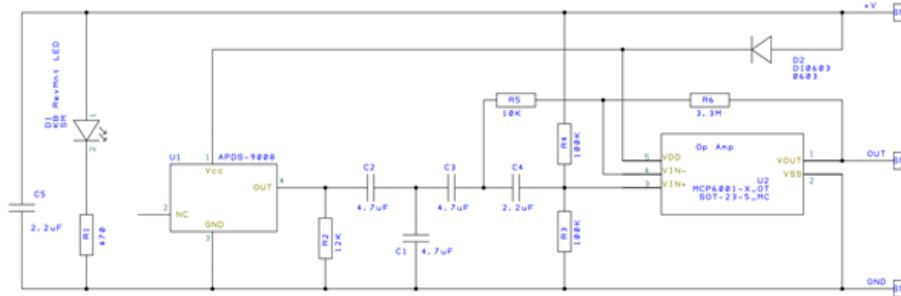


Fig. 2. Schematic sensor SEN-11574.

To condition the signal obtained from this sensor, the Microchip MCP3201 analog-to-digital converter (ADC) is used. This ADC has a sampling frequency of 50 kHz and a 12-bit resolution, sufficient to comply with the CENETEC standard, which indicates a minimum sampling frequency of 500 Hz and a minimum of 12-bit resolution for a basic electrocardiograph [10]. In addition, it has SPI (Serial Peripheral Interface) interface; in this way the digitized signal will be sent to the programmable device for processing.

High speed and sensitive response of system are necessary in measuring vital signs parameter. Measures can change drastically in a very short time, depending on the health status of patient. Well-known as high-speed IC processor, FPGA can work for specific purpose with high speed requirement [11,12]. Because of its flexibility and advantages, FPGA is perfectly suitable in medical instrumentation environment. The programmable device that will process the signals is a FPGA XC6LX16CS324 of the Spartan 6 family, implemented in a Xilinx Nexys 3 development circuit.

## 2.4 System Design

The purpose of the proposed system is to help physicians better monitor their patients. Considering that the system only contemplates the level of prototype, laboratory tests are carried out, placing sensors of the vital signs in students and professors of the Superior School of Computer Sciences (ESCOM). The overall process of the proposed architecture is shown in Fig. 3.

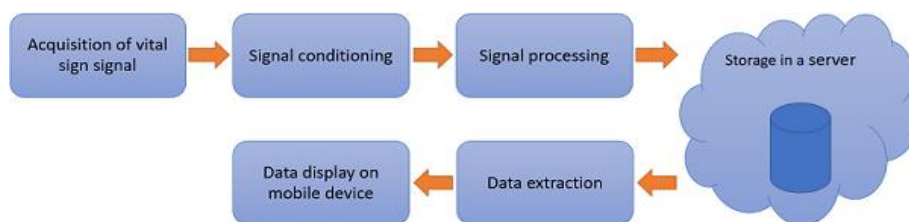
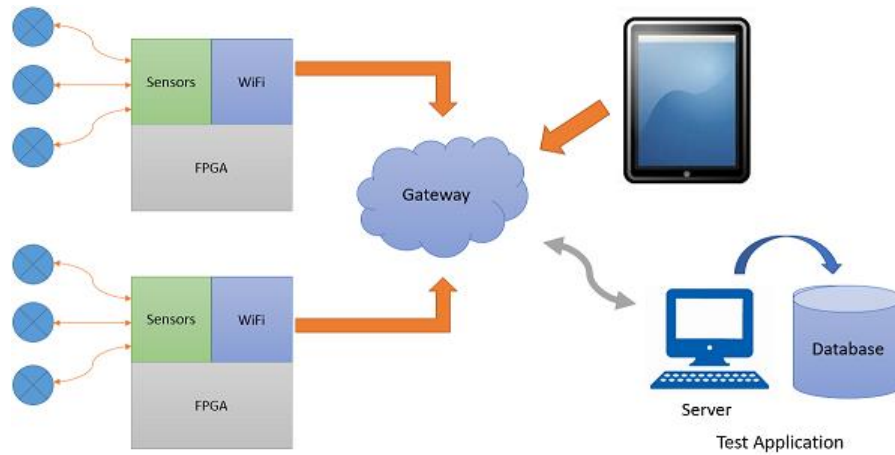


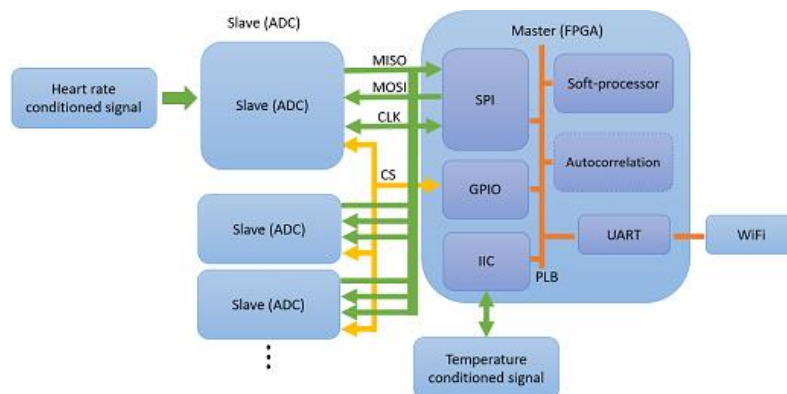
Fig. 3. Block diagram of the overall process.

The general architecture of mobile computing is shown in Fig. 4, which shows the way to communicate the sensor node with the server via WiFi. The mobile device connects to the server also via WiFi.



**Fig. 4.** General architecture mobile computing.

Fig. 5 illustrates the sensor node architecture in detail, it shows the master (FPGA) – slave (ADC) configuration via the SPI modules connected to the processor local bus (PLB). In addition, the IIC core to communicate with the temperature sensor is observed. The data obtained from the vital signs sensors are processed by the soft-core MicroBlaze. At the end of processing, the UART communication module connected to the PLB is used to send the results via WiFi to the server, using the WiFi ESP click module.

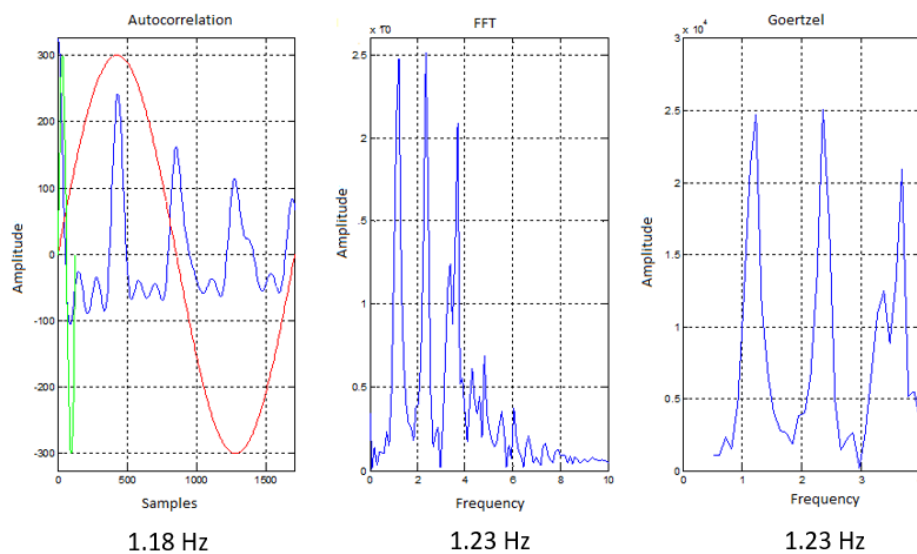


**Fig. 5.** Sensor node architecture.

## 2.5 Tests Performed

Initially, tests were performed on a sample obtained with a photoplethysmograph. The signal was digitized with a sampling frequency of 500 Hz and an 8-bit resolution. It was processed with different mathematical methods to obtain the fundamental frequency, such as: autocorrelation [13], fast Fourier transform and Goertzel algorithm [14].

**Choice of mathematical method:** The functions mentioned above can be implemented in software, running a program on a common computer or using a microprocessor or microcontroller based system. In this case, its performance depends on the computing power of the processor used and the processing speed which is usually medium in size, making it impossible to perform a real-time processing of high frequency signals. Another method is its implementation on FPGAs, which, due to their flexibility, possibility of parallel implementation and high integration capacity, have become the ideal hardware for those designs that are robust and require high speed [15]. After performing the calculations shown by MATLAB software, to choose the most convenient method, the graphs obtained are shown in Fig. 6.



**Fig. 6.** Comparison of mathematical methods for obtaining fundamental frequency.

The calculation of the fundamental frequency with the autocorrelation method is 1.18 Hz. This value is like that obtained with the Fast Fourier Transform (FFT) and the Goertzel algorithm, which is 1.23 Hz. Observing that the values obtained are similar, it can be considered acceptable any of the three methods. When implementing the algorithms on the FPGA, the computational cost would be higher using the Fast Fourier Transform, since the signal must be converted from time to frequency, and then find the maximum that indicates the heart rate. This process can be simplified by using the Goertzel algorithm, which is a frequency filter and does not perform the complete signal conversion in time, only uses a specific frequency range. With autocorrelation, the computational cost is reduced, since it is not necessary to transform the signal in time to frequency because the calculation of the fundamental frequency is obtained from the maximum value of the signal in a specific range of the samples.

The signal obtained with the photoplethysmograph was autocorrelated to obtain its fundamental frequency, obtaining the value of 1.18 Hz. After that, a Hamming window was applied to the original signal to avoid the Gibbs effect in frequency. When the windowed signal was autocorrelated, smaller values were obtained in the amplitude,

but the value of the fundamental frequency is the same. The graph of the windowed signal and its autocorrelation are shown in Fig. 7.

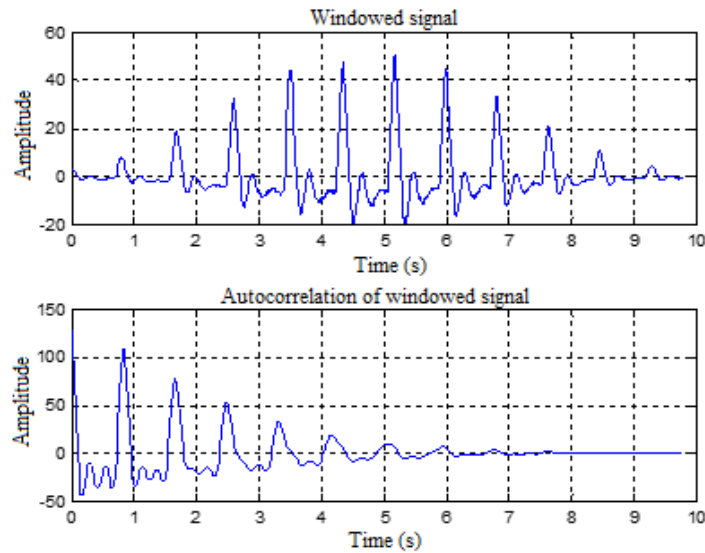


Fig. 7. Photoplethysmograph windowed signal with autocorrelation.

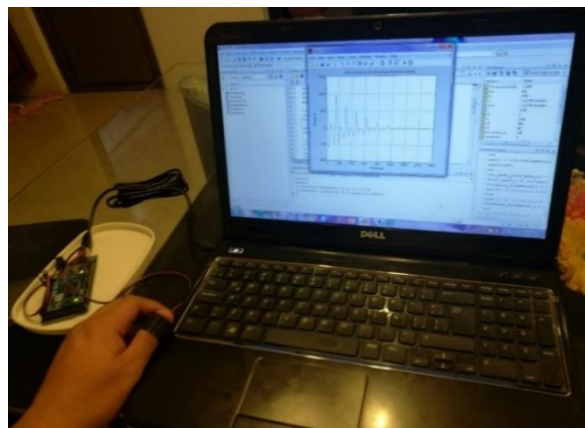
Table 2. Relative error percentage

Time (s)	1	2	3	4	5	6	7	8	9	10	
Number of samples	512	1024	1536	2048	2560	3072	3584	4096	4608	5120	
Frequency (Hz)	1.00	300.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Error (%)
	1.17	9.71	1.12	0.26	0.42	0.20	0.26	0.03	0.03	0.26	
	1.33	4.35	1.05	0.26	0.26	0.00	0.26	0.00	0.00	0.26	
	1.50	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	
	1.67	3.09	0.39	0.26	0.07	0.07	0.26	0.26	0.26	0.26	

**Calculation of necessary samples by measurement:** After testing with the signals obtained and validating the autocorrelation method for calculating the fundamental frequency, a script is performed to check how many seconds are necessary to have for a sample. Assuming ADC at 512 Hz sampling rate is used, we obtain the Table 2, which represents the percentage relative error in each case. Different signals are digitally generated using frequencies within the range of heart rate stability (between 60 and 100 beats per minute). It can be seen that among fewer samples are used, the error, compared to the real frequency value, is higher; from 7 or 8 seconds (3584 or

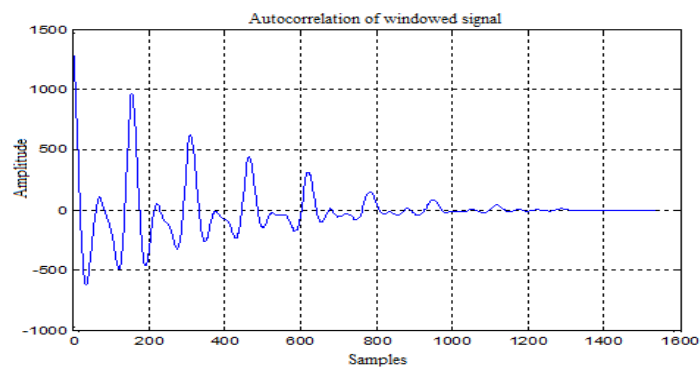
4096 samples) the error does not decrease significantly. Therefore, 4096 samples will be measured using the ADC at 512 Hz sampling rate.

**Communication photoplethysmograph - computer.** Once the number of samples needed to take measurements and the mathematical method applied to the signals have been defined, tests are performed with the chosen photoplethysmograph. Through a 192 Hz sampling rate and 10-bit resolution ADC, digitized photoplethysmograph signal is obtained. It is sent to a script in MATLAB by serial communication to perform the calculation of the heart rate (see Fig. 8).



**Fig. 8.** Communication photoplethysmograph with MATLAB.

The procedure performed after obtaining the signal is the same as that performed with the above photoplethysmograph signal. Fig. 9 shows the graph of the autocorrelation of the windowed signal; the result of the script indicates that the fundamental frequency of the signal is 1.23 Hz, ie 73.85 beats per minute (see Fig. 10).



**Fig. 9.** Autocorrelation of the windowed signal.

```
comienza
termina
La frecuencia fundamental de va es 1.23 Hz
La frecuencia cardiaca es 73.85 pulsos por minuto
```

**Fig. 10.** Result of the script on MATLAB.

### 3 Conclusions

In addition to the fundamental frequency results, the signal obtained by using electrodes to detect heart beats is more susceptible to noise than when using a photoplethysmograph. Since these sensors deliver an analog signal, it must be conditioned with an ADC that complies the medical standards of the CENETEC. The ADC must also be able to communicate with the FPGA, in this case, by SPI. Some sensors like the one used for temperature measuring might have digital outputs. Anyway, an interface is needed to communicate the sensor with the programmable device; in this case, this sensor has an IIC interface.

There are several advantages using an FPGA for the processing of this type of signals; for example, the ability to implement an HDL processor that is not limited to components or ports, plus the flexibility to add or remove components by reprogramming the FPGA and its ability to work in parallel.

Autocorrelation is very useful to find repetitive patterns within a signal, the periodicity of a masked signal under the noise and to identify the fundamental frequency of a signal that does not contain this component, but numerous harmonic frequencies appear from this one. This function has not been used in systems for obtaining vital signs that have been reviewed in this research, since all use the threshold method or that of obtaining the immediate frequency. Nor have we found systems that work with faster communication such as Ethernet or WiFi, which would allow a monitoring of the chosen vital signs in real time.

### References

1. Norma Oficial Mexicana del Expediente Clínico. Hojas de enfermería (2003)
2. Torres, M.: Calidad de los registros clínicos de enfermería: Elaboración de un instrumento para su evaluación, pp. 17–25 (2011)
3. Presbyterian, N.Y.: Vital signs: body temperature, pulse rate, respiration rate, blood pressure.
4. Neuman, M.: Measurement of Vital Signs: Temperature, pp. 40–49 (2010)
5. Neuman, M.: Vital signs: Heart Rate, pp. 51–55 (2010)
6. Neuman, M.: Measurement of Vital Signs: Breathing Rate and Pattern, pp. 39–44 (2011)
7. Neuman, M.: Measurement of Blood Pressure, pp. 39–44 (2011)
8. Bastar, S. G.: Metodología de la investigación, pp. 15–16 (2012)
9. Pérez, A.: Una metodología para el desarrollo de hardware y software embebidos en sistemas críticos de seguridad, pp. 70–75
10. CENETEC-Salud Cédula de Especificaciones Técnicas (2009)
11. Chu, P.: FPGA Prototyping by VHDL Examples. Wiley (2008)

*Jorge Martínez, Víctor García, Rubén Ortega*

12. Wakerly, J.: Digital Design: Principles & Practice. Prentice Hall (1999)
13. Vaseghi, S.: Advanced Digital Signal Processing and Noise Reduction, pp. 58 (2000)
14. Cortés, J.: Alternativa al análisis en frecuencia de la FFT mediante el algoritmo Goertzel, pp. 217–222 (2010)
15. Rice, J.: Configurable hardware solutions for computing autocorrelation coefficients: A case study (2005)

# Obfuscated Information Classification

González Rodríguez Florencio Javier, Aguirre Anaya Eleazar, Salinas Rosales Moisés,  
Barrón Fernández Ricardo

Instituto Politécnico Nacional, Centro de investigación en Computación, Mexico City, Mexico  
fjgonzalezr92@gmail.com, {eaguirre, rbarron }@cic.ipn.mx  
msalinasr@ipn.mx

**Abstract.** When a security assessment is executed, time and resources are limited therefore it is essential to identify those points that are most likely vulnerable and focus on those. Such points are identified in the information gathering stage turning it in a vital step during the assessment. Emerging a question about if gathered information could be manipulated by a third party either by functionality or by obscuring information goals. This paper proposes an information classifier in order to identify obfuscated information, classifying into obfuscated or integral information, with the purpose to be used during a “black box pentesting” assessment.

**Keywords:** black box pentesting, information gathering, security assessment, network fingerprinting.

## 1 Introduction

Information gathering is the first and most important phase in Penetration Testing. The goal is to obtain all possible target information with the purpose to get a security target profile like network architecture, features of devices involved, or personal user’s information that could be useful in the vulnerability testing. This kind of information should not be accessible for not authorized people, but in many cases it cannot be limited at all. For example, certain information in devices like open ports, protocol implementations, replay information from certain ports and so on. Penetration testers will take advantage of this to find vulnerabilities before attackers do.

Organizations used to mitigate such problem implementing some security controls like firewalls, letting block some patterns and limit access. Even with possible incongruence in configuration as is mentioned in [1]. On the other hand, for those information that cannot be limited at all, there are devices able to modify default protocol implementation fields, like protocol scrubbers that try to avoid protocol fingerprinting with the intention to mislead attackers.

This paper proposes a classifier for information collected from “information gathering” phase. Taking as input implementation features from the traffic generated during the information gathering execution in a black box pentesting, extracting default implementation values that can be changed either by the vendor or a security professional, in order to identify incongruities between network protocol implementations. Incongruities identification refers to compare basically results of three OS Fingerprinting techniques. First one analyzing TCP and IP protocol in order

to fingerprint the Operating System. Second one using ICMP protocol characteristics and third one analyzing some services banners results to identify modifications in the strings and also to fingerprint OS. Classification for each technique is executed using Machine Learning algorithms described in Section 5 “Analysis and Design”.

## **2 Security Controls for OS Fingerprinting Prevention**

Actually does not exist any tool that offers a target profile considering obfuscated information during a security assessment. This research proposes matching results through OS Fingerprinting using three techniques. Such techniques are related to some previous researches but having a different perspective. For example, can be found proposals like in [2] where Jason Barnes and Crowley make OS fingerprinting using a passive traffic fingerprinting mechanism, to identify hosts features involved in communications without interfering in any way. Using essentially SYN TCP/IP flags, obtaining features to identify HTTP clients, physical link types, and even if a host is behind a NAT device on a large network, receiving traffic from a Passive Network Appliance “PNA” in [3]. Being not necessary to make any system call, they have made evaluations in two laboratories: with constructed traffic and in an operational setting with real world traffic.

They compare their proposal with p0f and k-p0f tools, measuring the average maximum sustainable throughput across 30 second intervals 10 times for each mixture of traffic and type of monitor, resulting k-p0f better than p0f.

Same idea is analyzed in this proposal to fingerprint Operating Systems in order to compare results between different techniques and evaluate possibility of obfuscated information if any incongruities are identified.

In [10] Prigent, Vichot and Harrouet present IpMorph in order to show that fingerprint concealment and spoofing are uniformly possible against different known fingerprinting tools, IpMorph is a counter-recognition software implemented as a user-mode TCP/IP stack, ensuring session monitoring and on the fly packets re-writing used against fingerprinting tools like Nmap, Xprobe2, Ring2, SinFP and p0f. IpMorph cover more characteristics and analyses deeply OS Fingerprinters, even mention those able to get services banners in order to identify an Operating System, but they did not cover such aspect.

Our proposal covers the identification of those devices that protect an OS through the manipulation of banners structure. Therefore, is proposed a technique to identify if a Service Banner could have been manipulated.

Protocol scrubbers modify default fields of multiple protocols in order to reduce the number of techniques than can be used to identify an Operating System. Then they were analyzed to take in consideration protocols fields that usually protocol scrubbers modify. Analyzing congruency between protocols implementation and services banners.

### 3 Methodology

Experimental methodology was used in this research, since there exists a correlation between variables described in Tables 2 and 4 in Section 4, that were needed to analyze them through certain active and passive experiments, having results that are compared with expected ones, and take most descriptive values. Such methodology was divided in two phases:

**Exploratory:** This phase identifies questions that were tried to answer in testing phase, in this research such questions are mainly 3:

1. Can be identified obfuscated information?
2. Can be identified a device that obfuscates information?
3. Which features make identifiable obfuscated information?

**Testing:** In the testing phase planted questions were tried to be solved through experiments where analyzed values that could have been manipulated.

*Question 1* was answered through the analysis and comparative between default values for each device analyzed (NAT, Protocol Scrubber, Hardened host).

*Questions 2* were answered through the analysis of data type and behavior from values obtained during the analysis for each device.

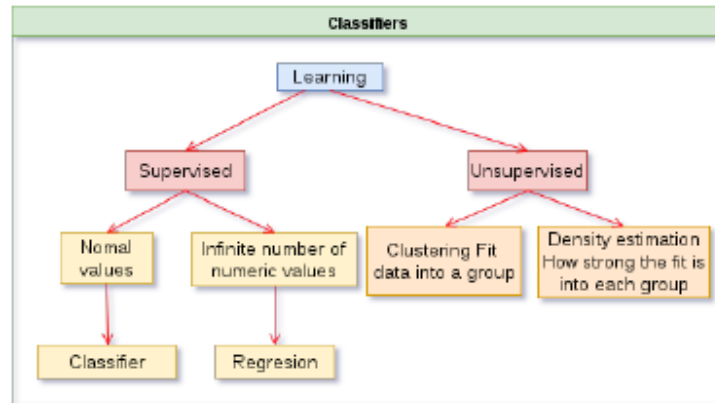
*Questions 3* were answered through the analysis of data type from values obtained during the analysis for each device, then testing Machine Learning algorithms and feature selection step.

### 4 Analysis and Design

#### 4.1 Algorithms

Classification is a way to solve a categorization problem using Machine Learning, there exist different algorithms used for solving certain problems shown in Fig. 1.

Machine Learning was used in this research to classify an Operating System given some features. Mainly are considered three elements for the classifier input, TCP + IP characteristics implementation, ICMP characteristics implementation and Service Banners analysis. First step is to identify the Operating System through OS Fingerprinting, analyzing TCP and IP headers of packets, due TCP/IP is the most common implemented protocol and most usable, besides TCP and IP have many field values that are not specified formally in RFC, it is known that certain security policies involve to block ports that are not used by a host, but it is possible even with blocked port analyze their behavior.



**Fig. 1.** Top Machine learning algorithms.

Mentioned inputs were selected analyzing related works about Protocol Scrubbers and NAT network analysis presented in Section 2. Characteristics for each input were obtained through an experimental process. First analyzing documentation about network Kernel parameters, Open Source Protocol scrubbers documentation and features analyzed in research papers. Then analyzing real traffic samples on different environments, in order to be compared versus documentation. Finally, such features are presented in subsection 4.3 for each implemented algorithm.

Mainly 3 open source protocol scrubbers were analyzed due others published protocol scrubbers are not open source, and it is not possible to analyze at all. Protocol Scrubbers analyzed were:

- IP Personality in [5]
- Scrub tech in [6]
- IP log in [7]

For the analysis in this research each Protocol Scrubber was installed in a virtual and physical machine. Studying configuration for each of them in order to get a list of values they use to change. As result were found 14 values that are used to identify Protocol Scrubbers. Being mainly TCP, IP, UDP and ICMP protocols those that are modified by Protocol Scrubbers, in order to affect indirectly things like RTT, timers and so on. Even packet length is used to hide an operating system implementation, due different payloads data used by each developer, in order to mislead a fingerprinting process executed by a tester. These variables are shown in Table 1.

It is possible to tag these features as an Operating System name, in order to identify an Operating System given some features.

**Table 1.** Values to identify Protocol Scrubber.

TCP	TS option	Urgent pointer	Winsize	Sack	Ack retries
	Nop option	Winscale	Max window	Options order	
IP	TTL	ID	ToS	Flags	
ICMP	Length	Payload			

In this research feature selection was made through a specialist person knowledge, letting to tag features for some systems. The type of features analyzed in each protocol are nominal values, therefore four classifier algorithms were selected to identify the OS, finally AdaBoost is implemented to choose best decision between classifiers.

Classifiers implemented for OS Fingerprinting using TCP+IP, ICMP and Services banners are:

- KNN
- Bayes
- Decision Tree

Classifiers were selected based on accuracy, training time, linearity, number of parameters and number of features. Taking in consideration data used in this proposal.

Each classifier was implemented in python using each algorithm model. Adapting parameters described in “Implemented algorithms” section, in order to get good results for classification. However, each of them was compared with Scikit framework, obtaining better results for some classifiers using the developed classifier and sometimes using Scikit. Best results are present through this paper.

## 4.2 Features

Features that were used by each algorithm are presented in section 4.3, however data used for training algorithms have a common structure shown in Table 2a, where  $x_n$  is a feature from the protocol and OS is the Operating System that is a label for features row.  $x_n$  is a field implementation protocol, or a value that can be obtained from interactions with the target. These values characterize or represent to the target, and can be used to make OS Fingerprinting.

**Table 2.** Features structure.

(a) Data structure used for classifiers training					(b) TCP+IP Data values used for classifiers training							
feature $x_1$	feature $x_2$	...	feature $x_n$	OS1	TTL	ToS	IP	Flags	WinSize	Sack	NopOption	linux
feature $x_1$	feature $x_2$	...	feature $x_n$	OS2	TTL	ToS	IP	Flags	WinSize	Sack	NopOption	windows
...	...	...	...	...	...	...	...	...	...	...	...	...
feature $x_1$	feature $x_2$	...	feature $x_n$	OS3	TTL	ToS	IP	Flags	WinSize	Sack	NopOption	linux

(c) ICMP Data values used for classifiers training				(d) ICMP Data values used for classifiers training			
Length	[Payload]	linux		linux	service	version	operating system
Length	[Payload]	windows		windows	service	version	operating system
...	...	...		...	...	...	...
Length	[Payload]	linux		linux	service code	service	version

The dataset used to classify using TCP+IP fields is shown in Table 2b and has the same structure shown in Table 2a. Such values are result of a previous analysis over Fingerprinting tools and those tools that try to avoid fingerprinting like Protocol Scrubbers.

The dataset used to classify using ICMP fields is shown in Table 2c and has the same structure shown in Table 2a.

Finally, the dataset used to classify using Services Banners is shown in Table 2d and has the same structure shown in Table 2a, but starting with operating system due strings longitude is not the same for all banners.

Data used for training and testing are samples shown in Tables 3 from Open source fingerprinting tools, different implementations extraction, and for Banners were used data from servers extracted using Shodan service. Each sample for TCP+IP and ICMP is a list of values for each field described in Table 2b, 2c with an associated label that is the Operating System name.

**Table 3:** Number of samples for training and testing.

Type	Traffic samples		Total
	# Sample		
	Physical	Virtual	
TCP+IP	45	8	53
ICMP	20	10	30
Banners (HTTP, FTP, SSH)	-	500	500

### 4.3 Implemented Algorithms

Three algorithms are evaluated using metrics in Equations 1, 2, 3 based on a confusion matrix for each algorithm:

$$Recovery(OS) = A_{ii} / \sum_{j=1}^n A_{ij}, \quad (1)$$

$$Precision(OS) = A_{ii} / \sum_{j=1}^n A_{ji}, \quad (2)$$

$$Accuracy = \sum_{i=1}^n A_{ii} / \sum_{i=1}^n \sum_{j=1}^n A_{ij}, \quad (3)$$

where Recovery is the proportion of cases correctly identified as belonging to class C among all cases that truly belong to class C. Precision also called true positive rate, is the proportion of cases correctly identified as belonging to class C among all cases of which the classifier claims that they belong to class C. Finally, Accuracy is the ratio of correct predictions to total predictions made.

### Naive Bayes

Naive Bayes is a probabilistic classifier based on the Bayes theorem with strong naive independence assumptions between the features. Due this classifier assume that the value of a particular feature is independent of the value of any other feature, given a class C. For this proposal it is important because protocol scrubbers modify values from time to time, then if values were dependent, classifier will not work at all as Decision Tree classifier that will be also described. Equations used in this research for Naive Bayes are show in equation 4:

$$P(C|x_1, x_2, \dots, x_n) = \frac{(\prod_{i=1}^n P(X_i|C)P(C))}{P(x_1, x_2, \dots, x_n)} = \frac{P(x_1, x_2, \dots, x_n|C)P(C)}{P(x_1, x_2, \dots, x_n)}, \quad (4)$$

where:

$$P(C) = \frac{\text{Number of } C \text{ classes}}{\text{Total number of classes}},$$

$$P(x_n|C) = \frac{\text{Number of rows that have } x, \text{ and are } C \text{ class}}{\text{Number of rows that are } C \text{ class}},$$

$$P(x_1, x_2, \dots, x_n|C) = P(x_1|C)P(x_2|C) \dots P(x_n|C),$$

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2) \dots P(x_n).$$

Table 4 shows the confusion matrix for Naive Bayes, with 20 samples for each Operating System where 10 were taken from training samples and 10 were different from training samples.

**Table 4.** Confusion Matrix for Naive Bayes.

Windows 9x/NT 20	OSX 10.x	Linux 2.2.x	Linux 4.x	Cisco 12.0	OpenBSD 2.x	Windows 9x/NT 20
Linux 2.2.x		18	2			
Linux 4.x		1	19			
Cisco 12.0				20		
OpenBSD 2.x		1			19	

Naive Bayes Classifier evaluation

**Table 5.** Evaluations metrics for Naive Bayes.

	Recovery Precision	
OS		
Windows 9x/NT	1	1
OSX 10.x	0.9	0.9473
Linux 2.2.x	0.9	0.9473
Linux 4.x	0.95	0.9047
Cisco 12.0	1	1
OpenBSD 2.x	0.95	0.9047

$$\text{Accuracy} = 114/120 = 0.95$$

### K nearest neighbours

K nearest neighbors is a classifier that stores all available cases and classifies new cases based on a similarity measure. Equations used in this research are show in Equation 5:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}, \quad (5)$$

where:

$x_i$  is the feature in the database

$y_i$  is the input feature to classify

$d$  is the distances that means how different are the input and the database item

Table 6 shows the confusion matrix for KNN, with 20 samples for each Operating System where 10 were taken from training samples and 10 were different from training samples.

**Table 6.** Confusion Matrix for KNN.

Windows 9x/NT	OSX 10.x	Linux 2.2.x	Linux 4.x	Cisco 12.0	OpenBSD 2.x
Windows 9x/NT 20					
OSX 10.x	17				3
Linux 2.2.x		17	3		
Linux 4.x		2	16		2
Cisco 12.0				20	
OpenBSD 2.x	3				17

KNN Classifier Evaluation Metrics obtained from confusion matrix in Table 7 are calculated as: Accuracy =  $107/120 = 0.8916$

**Table 7.** Evaluations metrics for KNN.

	Recovery	Precision
OS		
Windows 9x/NT	1	1
OSX 10.x	0.85	0.85
Linux 2.2.x	0.85	0.8947
Linux 4.x	0.8	0.8421
Cisco 12.0	1	1
OpenBSD 2.x	0.85	0.7727

## Decision tree

Decision tree is a predictive model where the target variable can take a discrete set of values, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

Table 8 shows the confusion matrix for Decision Tree, with 20 samples for each Operating System where 10 were taken from training samples and 10 were different from training samples.

**Table 8.** Confusion Matrix for Decision Tree.

Windows 9x/NT	OSX 10.x	Linux 2.2.x	Linux 4.x	Cisco 12.0	OpenBSD 2.x	Windows 9x
OSX 10.x	20					
Linux 2.2.x		17	3			
Linux 4.x			15			5
Cisco 12.0				20		
OpenBSD 2.x					10	10

Decision Tree Classifier Evaluation Metrics obtained from confusion matrix in Table 9 are calculated as: Accuracy =  $102/120 = 0.85$

It is worth to mention that Decision Tree had bad results when features do not exist in the database, however this property could be useful to identify specific hardened hosts or NATted networks in future work. Due hardened hosts just change their default implementation values just when is hardened and is not making changes repeatedly,

also it is known that exist some hardened hosts distributions. Such systems can be analyzed and added to the database, ensuring that is going to be identified by Decision Tree algorithm.

**Table 9.** Evaluations metrics for Decision Tree.

Recovery Precision		
OS		
Windows 9x/NT	1	1
OSX 10.x	1	1
Linux 2.2.x	0.85	1
Linux 4.x	0.75	0.8333
Cisco 12.0	1	1
OpenBSD 2.x	0.5	1

### ADA Boost

ADA Boost is an algorithm for constructing a “strong” classifier as a linear combination of others classifier referenced as “weak”. General idea is represented by Equation 6:

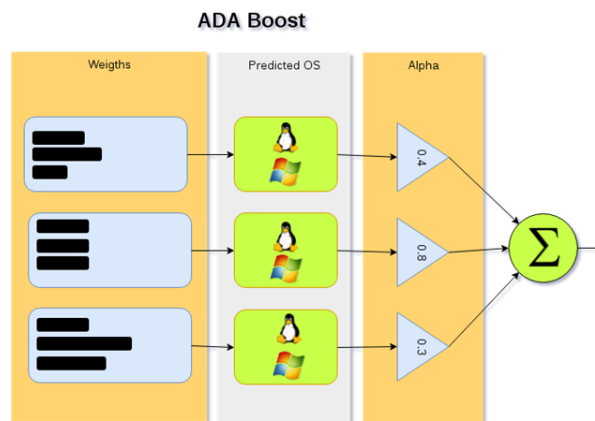
$$f(x) = \sum_{t=1}^T a_t h_t(x), \quad (6)$$

where:

$h_t(x)$  is a “weak” classifier

$\alpha$  is an assigned weight for each instance in the training dataset.

Each weighted prediction pass through a classifier, which is then weighted as “alpha values”. Finally, each alpha value is summed up in the circle that processes the final result as Fig. 2 shows.

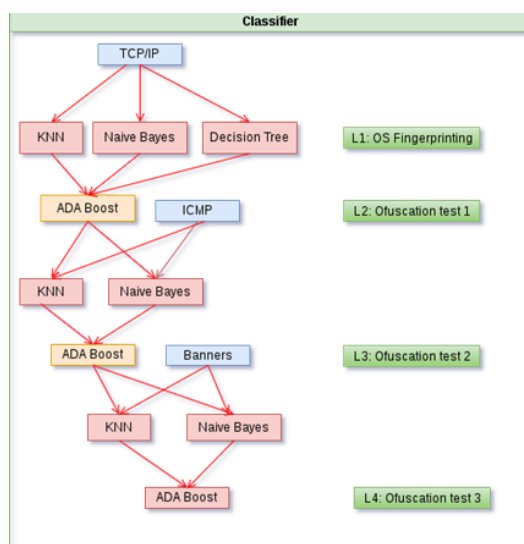


**Fig. 2.** Banners classification for OS Fingerprinting.

In this research weak classifiers are Naive Bayes, KNN and Decision Tree. Obtaining an Operating system and a weight after classification process. Resulting with a most representative Operating System for the IP address analyzed.

## Proposal

Fig. 3 shows the proposal based on results shown in Tables 5, 7 y 9. Where is executed an OS Fingerprinting based on TCP+IP, ICMP and Services Banners analysis. It is divided in three steps, first one is to identify an Operating System using TCP+IP characteristics using three tested classifiers, having as a result three Operating Systems that presents just one IP Address. If some of them are different then they could be manipulating information, that is considered as information obfuscation, decision that is taken by ADA Boost algorithm. Second step is the same idea but using as input ICMP characteristics, also considering last result in order to match and get a congruence value as reference. Finally, Banners are analyzed, considering previous results and having as a result an Operating system name and a value that represents congruence between results.



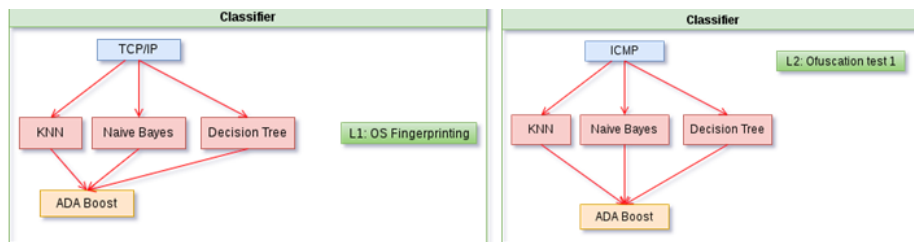
**Fig. 3.** Classifiers proposal.

Because of the number of features of TCP + IP useful for OS Fingerprinting, these protocols are analyzed to identify with most accuracy the Operating System. Each classifier has a result that should be the same for the three classifiers, because it means that network implementation for the host has not been modified, but if incongruities are identified then it could be obfuscating information.

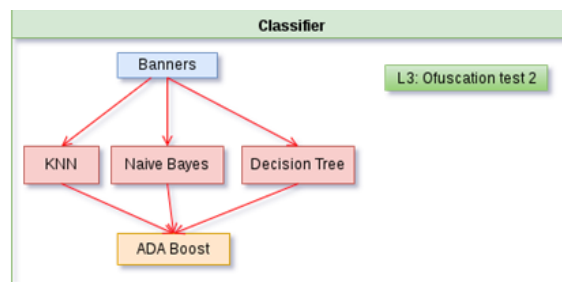
In order to have just one result ADA Boost is implemented to choose the best decision choosing an OS that identifies an IP address. Finally, ADA Boost has just one result as output in order to be analyzed in the next step as Fig. 4a shows.

For the next step is analyzed ICMP, that offers less features to identify an Operating System, but they are enough to detect ambiguities between TCP+IP and ICMP, also joining results for each classifier through ADA Boost as is shown in Fig. 4b.

Finally, Services Banners analysis are the last step due services that were installed after the initial hardening offers clues about the Operating System, obtaining a new feature to compare against TCP/IP and ICMP analysis, having as an output a numeric value that represent if exist ambiguity between the three analysis, letting to classify as obfuscated or integral information. Structure is shown in Fig. 4c



(a) TCP+IP classification for OS Fingerprinting. b) ICMP classification for OS Fingerprinting.



(c) Banners classification for OS Fingerprinting.

**Fig. 4.** Proposal classification algorithms structures for OS Fingerprinting.

## 5 Results Analysis

In Table 10 can be seen evaluation metrics for each algorithm obtained from previous reported tests.

For Windows 9x/NT, default values in the dataset are kind of different compared with other OS samples, range of values have big numeric differences versus others, so results are as expected, classifiers are able to identify Windows 9x/NT 100% of tests. In OSX 10.x percent is not the best for those algorithms that offers how similar is the input versus the training dataset. It is because of as it is known OSX is an Operating System based on BSD Family OS, then it can be seen in Tables 4, 6 that sometimes classifier confuses OSX with BSD. However, Decision Tree did not make any mistake, it is because Decision Tree classifies it just if all characteristics match with a dataset sample.

For Linux Operating system with different Kernels it is the same as OSX analysis. Classifiers confuses Kernel because of Operating Systems Kernels are similar. But if they have unique default values, they are classified correctly by Decision Tree, for example old kernel versions of Linux.

Finally, for other systems with no similar characteristics with dataset samples, for example Cisco 12.0 and Windows 9x/NT, are identified without any problem. But if more similar versions are added, then classifiers will start making mistakes. Even when classifiers made some mistakes, accuracy keeps an acceptable value. Results that are evaluated by ADA Boost algorithm in order to get best results from classifiers.

**Table 10.** Evaluation metrics for implemented Classifiers.

OS	Bayes Recovery Precision		KNN Recovery Precision		Decision Tree Recovery Precision	
Windows 9x/NT	1	1	1	1	1	1
OSX 10.x	0.9	0.9473	0.85	0.85	1	1
Linux 2.2.x	0.9	0.9473	0.85	0.8947	0.85	1
Linux 4.x	0.95	0.9047	0.8	0.8421	0.75	0.8333
Cisco 12.0	1	1	1	1	1	1
OpenBSD 2.x	0.95	0.9047	0.85	0.7727	0.5	1
Accuracy	0.95		0.8916		0.85	

## 6 Future Work

The current proposal represents the base for obfuscated information identification, but it is needed to identify as well those devices that are able to obscure integral information, therefore future work involves identification of devices like NAT Networks, Protocol Scrubbers and Hardened hosts. Having as a result obfuscated information but also those devices able to obfuscate them. Letting to a tester identify these kind of security controls, concentrating on systems emanating real information that can be used to compromise them.

## 7 Conclusions

Obfuscated information identification hypothesis through protocol implementation analysis is demonstrated using classification algorithms, even with features obtained from passive information gathering.

Each algorithm has advantages and disadvantages as can be seen in result analysis section, then Adaboost complements the work to get best results for each test. Decision Tree and KNN seems to be worst algorithms to classify this kind of problem. However, it is not in this research, because of if a specific obfuscation system was identified and integrated to data base, it has high probability to be identified using this proposal. For example, a specific version of a protocol scrubber, a public hardened host system and so on. For Bayes can be seen that have best results. Due Bayes is probabilistic, letting

modify some equation properties in order algorithm decide based on a binary value, but having a result about how much is similar the input with database samples.

Then based on results it is possible to identify obfuscated information taking advantage of some network protocols implementations. Just analyzing any incongruence in their configurations between those modified protocols. Task that is valuable for penetration testers during a black box security assessment, due it is important to focus over those most likely vulnerable systems in the environment, reducing time and resources. But also having present those possible devices that could be a security control.

**Acknowledgment.** The authors gratefully acknowledge use of the services and facilities of the Centro de Investigación en Computación (CIC) and Instituto Politécnico Nacional (IPN), also to Consejo Nacional de Ciencia y Tecnología (CONACYT) for supporting this research.

## References

1. Antonis, P., Polydoros, P., Miltos, G.: A firewall module resolving rules consistency. TEI of Crete (2017)
2. Jason, B., Patrick, C.: k-p0f: A High-Throughput Kernel Passive OS Fingerprinter. Washington University in St. Louis (2013)
3. Schultz, M., Ben, W., Patrick, C.: A Passive Network Appliance for Real-Time Network Monitoring. Washington University in Saint Louis (2011)
4. Guillaume, P., Florian, V., Fabrice, H.: IpMorph: fingerprinting spoofing unification. Plouzan, France (2009)
5. IP Personality: <http://ippersonality.sourceforge.net>
6. Scrub tech: <http://scrub-tech.sourceforge.net/>
7. IPLog: <http://ojnk.sourceforge.net/stuff/iplog>
8. Massimiliano, A., Ermanno, B., Sushil, J.: A deception based approach for defeating OS and Service Fingerprinting. Naples, Italy (2015)
9. Matthew, S., Robert, M., Farnam, J.: Defeating TCP/IP Stack Fingerprinting. Ann Arbor, Michigan (2000)
10. Steven, M.: A technique for counting NATted Hosts. Marseille, France (2002)
11. Liang, W., Kevin, P., Aditya, A., Thomas, R., Thomas, S.: Seeing through Network Protocol Obfuscation. Denver Colorado, USA (2015)
12. Quinlan, J.: Induction of Decision Trees. Sydney, Australia (2017)



# Mobile Computing System for Neuropsychological Evaluation

Enrique Alfonso Carmona García<sup>1</sup>, Elena Fabiola Ruiz Ledesma<sup>1</sup>,  
Laura Ivoone Garay Jiménez<sup>2</sup>, Mario Eduardo Rivero Ángeles<sup>3</sup>

<sup>1</sup> Instituto Politécnico Nacional, ESCOM, CDMX, México  
eacarmona860920@gmail.com, efruiz@ipn.mx

<sup>2</sup> Instituto Politécnico Nacional, UPIITA, CDMX, México  
lgaray@ipn.mx

<sup>3</sup> Instituto Politécnico Nacional, CIC, CDMX, México  
mriveroa@ipn.mx

**Abstract.** The cognitive process evaluation requires the application of a neuropsychological test such as the *Evaluación Neuropsicológica Breve en Español (Neuropsi ©)*. It is a standardized and validated test in Mexico. This test consists of a battery of activities aimed at measuring the memory, attention, motor, language, and vision-space skills, considering the age and education level. Nowadays, the application is performed by a specialist and he must observe, measure and take note of the process for the final evaluation, this procedure takes more than one hour per subject. Considering this test as a possible evaluation for the student cognitive skills, this period is too long for application in a class in real time. So, it was proposed a software that implemented the Neuropsi © test, it could be carried out simultaneously by different users on its own mobile devices. The information is concentrated in a server and the system sends quantitative and qualitative automated evaluation to the specialist and he can obtain individual and group profiles. In the present paper, we report the modular architecture of the system according software engineering, the diagnostic and results modules and its functionality is described.

**Keywords:** neuropsychological test, automated cognitive skill evaluation.

## 1 Introduction

A current goal for the educational technology is to standardize the performance evaluation into the teaching-learning process but because of its complexity, it must be analyzed from an interdisciplinary perspective. In the present work, the main interest was to develop a modular architecture of a computer system according the software engineering. This system measures and promotes the attention and concentration simultaneously in a group of students, with the purpose of reporting the improvements for each student or for an entire class to a specialist.

There are several factors that cause low academic achievement, some of these factors are lack of understanding of the involved concepts of the studied topic, lack of good study habits or strategies, lack of attention in classes among others [1,2]. Implementation of pedagogical solutions that allow raising the quality of learning

processes through the reinforcement of cognitive abilities requires monitoring these processes and reinforce these basic abilities.

One of the most important elements is the attention, which is regulated by neurological centers [3-9]. Any biological or physical abnormal behaviors of them, alters the final result: to attend effectively to the respective task. Attention, as a psychological process, is characterized by the transformation of its basic reflected nature, such as the case of attention due to an imminent danger, by acquiring a voluntary character from the interaction with the medium [10], [11]. Concentration is considered as a prolonged process of attention and occurs when there is a strong motivation towards the tasks, activities or demands of the environment [11], [12]. The study of the attention and concentration characteristics by specialists has allowed to establish different levels of attention [10], [11]. The levels of attention are: focused, held, selective, alternating, and divided attention. They are measured by specially through a designed test for this purpose. It has been possible to establish the hierarchy that exists in these different kinds of attention. Noting that to be successful in complex tasks which requires high attention levels, such as alternating and divided attention, firstly, it is necessary to train the held and focused attention [10], [13]. In general, for the evaluation and measurement of attention, direct observation can be used [11], [12] and questionnaires such as the Conners scale [13]. In neuropsychology clinic, the accepted instruments of measurement are the neuropsychological tests, that evaluate a determined cognitive domain or skill, defined as a sample of conduct under specific conditions [14]. In the present study, we selected the neuropsychological brief evaluation battery in Spanish, Neuropsi ©. This neuropsychological battery is a standard collection of several tests for specific cognitive skills. The administration of the battery and obtaining the results usually takes a period which varies between days and weeks because the application of the battery is done in a personalized interview and the specialist is taking note of the relevant qualitative and quantitative characteristics in each test as well as the registration of the specific times associated to the performance of the subject during each test.

However, in a classroom where the students in the group have already been over certain academic requirements for admission, and they already are considered as typical healthy condition, it is important to evaluate, in a fast and reliable way, the present conditions of their attention capacities. This test could be used to have feedback into the teaching-learning process and to alert about a dangerous diminish of students' performance.

The paper is organized as follows: Section II presents a review of the existing literature about attention and concentration and the actual tests and batteries that evaluate these processes. Section III presents the methodology proposed for the development of the modular architecture of the computer system. Finally, section IV presents the preliminary results and part of the future work.

## **2 Related Work**

The first studies about attention and concentration appears at the end of the 19th century and the beginning of the 20th [15]. At present, many of the accepted concepts and ideas are dated from that beginning studies and others realized between the years 50s and 70s

of the last century [13], [16], [17], [18]. Attention is the ability to address some specific stimuli, inhibiting other simultaneous external or internal stimuli [19]. Attention and concentration have several parameters that let analyze their practical use in different contexts. These characteristics are: selectivity, volume, cyclicity, direction, intensity, and stability [11].

Evaluation of attention and concentration is often a complex process mainly due to the nature of the processes itself and its dependence of the context of the evaluation. In order to select the tests to be used into the proposed system, it should be considered the specific condition of application and it is necessary to meet psychometric criteria for reliability and validity into the description of the abilities and cognitive disabilities of an individual [20], [21]. It is important to pay attention on the educational level and the age of the subjects to make an adequate interpretation of the results. Also, it has been proven that the tests application to non-native speakers or people with different values or culture shows up significative difference in the obtained results [10], [22]. Because of this, although many functional tests and evaluation scales are available in English, it is not convenient to use them in Spanish-speaking and / or low-income and / or low-culture populations [20]. To solve this issue, several diagnostic tests for assessment of attention and concentration had been adapted so they can be used in several application cases. Some of these tests are: digit memorizing [22], [23], [24], *Corsi cubes* [18], *continuous cancellation or execution tests* [22], [25], *symbolic digit test* [26], *paced auditory serial Addition test* [27], *trace test* [25]. It is important mention that all these tests are independent of each other and they are specialized in assessing one or more specific levels of attention and concentration. Several batteries with specific purpose had been reported [22], some of these evaluation batteries are: *Halstead-Reitan*, *Luria-Nebraska*, *Ardila-Ostrosky*, *Dementia Rating Scale*, *Mental Status Check List*, *Blessed Orientation-Memory-Concentration test* [8], [9], [12]

A neuropsychological examination with a specific battery, like any type of clinical and paraclinical examination, has tests and applies procedures to achieve specific objectives [28]. Nowadays, there may be many reasons for conducting a neuropsychological evaluation such as determining work ability, following up on a psychological treatment and being able to determine the characteristics of its evolution or simply for exploratory or research purposes.

Specialist usually selects a neuropsychological battery to do the initial diagnosis of the patient and once the results have been analyzed and the diagnosis performed, a new specific test is selected to entrain the patient in specific skills. Nowadays these processes of selection, application of the neuropsychological battery, evaluation of the results and selection of complementary tests, are carried out by the specialists in a manual way. There are some specific tests that can be found in digital format for application or some basic evaluation tools, but they are not frequently used by the neurologists because administering and evaluating some tests in digital format and another in a manual way makes difficult to manage them in the cases of being include in a test battery. Besides, although in our days exist many different informatics solution who test some cognitive abilities like attention, visual perception, planning, reaction times, they are mainly oriented to children and to improving their general abilities. [29], [30], [31], [32]. In the case of Cognifit® [32], the system provides a complete computerized battery of tasks that allows the assessment of over 20 fundamental cognitive functions, clearly defined with objective measurement controls that provide

standardized age and demographic criteria based on the results of thousands of subjects. All the tests and tasks are validated by independent studies in representative populations, using the placebo system to measure and follow their effectiveness [33], [34]. In general, all this system present cognitive and neurological assessments. Unfortunately, the validation in Mexico is not reported and all these systems required to pay a license to be used, and in some cases their license is expensive.

For the implementation of the system, the brief Neuropsychologic battery in Spanish was selected, which it is validated for Spanish-speaking populations, it is ranked by level of education as 0, 5-10 and >10 years of education and age into a range 5 to 75 years old [19] [20]. This test battery measures important aspects related to the cognitive process, functions such as attention and concentration, spontaneous memory, visio-space functions, comprehension, executive and evocation functions.

### **3 Methodology and Partial Results**

The previous selection of the test battery was based on the need to develop an advance educational software that allows to administer and evaluate a neuropsychological battery for students in a website. Because the application of the selected neuropsychological evaluation starts with the selection of the battery to be used and ends with the final diagnosis generated by the specialist, the system is designed with the intention of reducing the time assigned to the administration of the battery helping to the specialist to manage an increasing amount of data.

The scores obtained in the different tests integrate the quantitative data considered by the specialist for the diagnosis. Each of these variables could have an independent effect or not [33], [34]. In the last case, there are some variables such as educational level, age, sex, culture, and laterality that should be considered [14]. The neuropsychological evaluation methodology has four stages: initial clinical evaluation, selection and application of the follow-up tests, analysis of results and generation of the report.

At present, there are a variety of software applications that are currently used in our daily lives, whether into our residence, traveling and working. In specific the mobile applications are used as a tool for a specific operation or complement a function in our modern life. One of the main reasons for creating a mobile application is the need to solve a problem or to simplify a complex operation in a ubiquitous way [35]. So an important concern in the software design is saving the generated information in an effective way so that it can be consulted or be analyzed at any time.

There are many database models, such as the relational models, non-relational models, hierarchical and object-oriented among others. The most significant challenge in implementing a database is the development of database structures [36], [37], [38]. It has been selected MongoDB for the storage of the generated data because MongoDB is a scalable, powerful, and flexible general-purpose database. It combines the ability to scale with features such as secondary indexes, rank queries, sorts, aggregations, and geospatial indexes. It is a database oriented to documents, replacing the concept of rows by another model of greater flexibility, documents; This means that, instead of storing the data in tables, MongoDB stores data structures with a specification called BSON, using a dynamic schema [39], [40].

In this case, two separate databases are used: User Database and Evaluation Database. In the database of users will be stored the information belonging to all people who use the system, regardless of its role in it. In addition, it will store the results obtained by the students when performing different batteries and tests. On the other hand, the Evaluation Database will contain all the information of the batteries and tests that the system has.

To successfully achieve the objectives of the system, it is necessary to use a development methodology capable of carry out the required processes efficiently. In this context, a process defines who is doing what, when, and how is performed a task. In software engineering, the aim is to build a software product or upgrade an existing one.

Following this methodology is guaranteed a high-quality software, reducing risk, and obtaining a predictable project [41]. The development process sets the needed activities to transform a user's requirements into a software system. The rational unified process (RUP) is a generic framework that can be adopted for a wide variety of software systems with different project sizes. RUP is based on software components interconnected through well-defined interfaces. The main aspects in RUP are: case-driven, centered into the architecture and iterative and incremental methodology [41], [42]. And it uses the Unified Modeling Language (UML) to prepare all schemas of a software system design for technical documentation [42], [43].

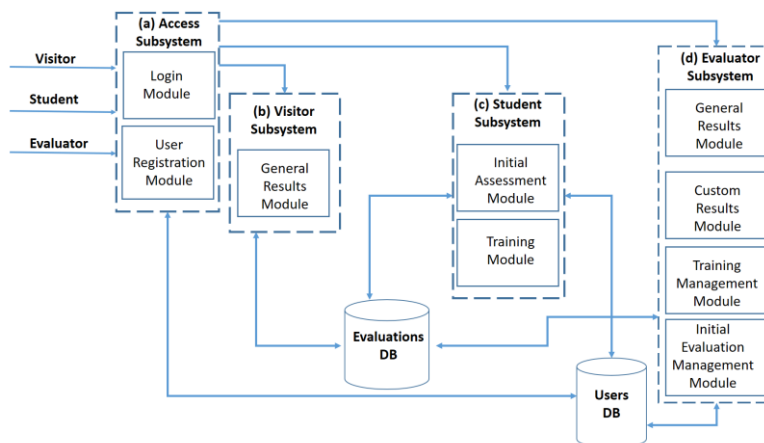
Because of the design of the system, the administration and evaluation of a single test or a tests battery can be carried out at any time and place so the student or the medical specialists do not need to be in the same place at the same time to applying the test, even it could be applied by a teacher in a classroom after a short training. Another important issue to be considered, it is the integrity, reliability, and security of all the information that is contained in the databases.

In Fig. 1 is presented the architecture of a modular system. It was designed to give independence between modules into the subsystems. It provides a solution considering an efficient maintenance. Another advantage is related to stability and availability of the final software because the administrator can disable a specific module and it does not affect the rest of the software so it continues available.

The Access Subsystem shown in Fig. 1 case (a) was composed by two modules: Login Module and the User Registration Module. In Login Module, it was implemented a role-based authentication to identify the potential users of the system. This module identifies the role of the user at the time he is login into the system, so all the not required system sections or information are maintained transparently and independently to the user. The administrator of the system directly supervises the User Registration Module because he is the person authorized to manage all user in the system so he adds, removes, or changes all the information regarding of the registered user in the system.

The heart of the architecture is based on the Student Subsystem and the Evaluator Subsystem, shown in Fig. 1 case (c) and case (d). The Student Subsystem is composed by two modules, the Initial Assessment Module, and Training Module. The main purpose of the Initial Assessment Module is provided the *Evaluación Neuropsicológica Breve en Español (Neuropsi ©)* to set the initial level of attention of the student and classify it as low, normal, or high according to the medical criteria. Afterward, the Training Module is activated. This module supplies tests to maintain or improve the level of attention. On the other hand, the Evaluator Subsystem manages all the test

offered by the system, the evaluation process and show the result to the specialist. Four modules compose the Evaluator Subsystem: General Results Module, Custom Results Module, Initial Evaluation Management Module and the Training Management Module.



**Fig. 1.** Mobile System Architecture for Neuropsychological Evaluation. The subsystems are a, b, c and d. Modules are into the subsystems.

The General Results Module evaluates the results of the applied tests and determine the levels of attention as low, normal or high according the medical standard. The Training Module is called if a suggested action is required to improve the cognitive skills. The Custom Results Module allows to present the result obtained in several formats such as tables, graphic and provides files compatible to generic tools to facilitate the interpretation and analysis.

It important to point out that the General Results Module of the Evaluator Subsystem and the Visitor Subsystem are the same module, the difference is in the available results and its format that is offered to a visitor and an evaluator. The visitor only can access to the class information without identification information associated to the students but the Specialist, in the evaluator role, could access to every result obtained. The Initial Evaluation Module is responsible for generating and providing the complete validated neuropsychological test (*Neuropsi* ©) that is currently active in the system for the new student users.

Finally, in the Training Management Module, an evaluator could remove or edit any of the test in the system and he also has the possibility to add new test. It is worth mentioning how this module implemented the new offered test. All considered tests from the validated battery were analyzed offline, to identify their logic, then they were implemented, modifying the difficulty level according the clinical criteria. This methodology provides a greater number of test without need to introduce them by the specialist one by one. The advantage is that each test retains measured variables, clinical skills levels considered and the application of constant parameters, so into the new test the evaluation procedure is maintained. It is important to mention that the specialist can create as many tests as it is required. Also, the system provides to the

evaluator role the option of creating new entire neuropsychological batteries, customizing the tests that are already in the system for research or new studies validation.

This increase of available resources represents a significant improvement compared to the current application method because with this system it can be generated as many neuropsychological batteries with their respective standard evaluation forms in an easy way to make variations to them, the specialist could define the order of application and the tests can be printed for application in a traditional way or be applied into the system. Another important point to highlight is that this digital implementation of the neuropsychological tests batteries gives an auxiliary tool to the specialist to applied the test simultaneously through the web application assisted by a teacher previously trained. Moreover, the automatically measurement of the variables considered test and automated evaluation is an improvement with respect to the traditional way, where the specialist must annotate, observe and measure response times, completion times, which impacts on the number of people he can attend and the time taken to apply the test.

## **4 Conclusions**

As far as the authors knows, the Mexican psychologist do not have a validated mobile application that allows them to add existing neuropsychological tests batteries for their digital application. Neither they have digital evaluation tools that help them during the evaluation process so this automated system could significantly reduce the time needed for managing and evaluating the considered neuropsychological battery. This paper presents a modular architecture of a computer system for the application and evaluation of entire batteries or specific neuropsychological tests focused on attention and concentration so it was described how the system was structured and the functionality of the different modules that are into the system. The software engineering goals considered was to minimize as far as possible the application time of generating a test and the evaluation time required for test application, to give ubiquity and to increase the temporal availability of the system. The main contributions of the proposed computer system are:

- 1) Automated application and evaluation of neuropsychological tests batteries.
- 2) Creating new tests maintaining the skill level, evaluation variables and the original application parameters.
- 3) Obtaining new performance information, which are not currently measured in the traditional way of batteries application, such as an interval of time between each answer.
- 4) Easy generation of additional tests of attention and concentration but gradually ramping up the difficulty.
- 5) Storage of the obtained information in a centralized way and using a non-relational database, making possible its use in future knowledge discovered.

Because the easy way to create, visualize and test new forms of evaluation, it will possible to carry out new neuropsychological studies that will be concentrated, update and modified online providing an invaluable tool for research about the cognitive process into classrooms.

**Acknowledgments.** The authors thank to Instituto Politécnico Nacional for the partially supports by grants SIP 20164801 and SIP20170137 and the scholarships provided to Enrique Alfonso Carmona García. A special recognition to Dr. Azucena Lozano for her valuable advice in the neuropsychological area and to Dr. Ostrosky-Solís for its authorization to use the Neuropsi © test as the validated test in Mexico.

## References

1. Ruiz Ledesma, E.F.: Sistema de evaluación de procesos cognitivos basado en estándares psicopedagógicos haciendo uso de tecnología educativa. Instituto Politécnico Nacional, Registro en la SIP 20164801 (2016)
2. Ruiz Ledesma, E.F.: Impacto del Cómputo Móvil en la enseñanza del concepto de variación. Instituto Politécnico Nacional, SIP 20131825 (2013)
3. Uriarte Bonilla, V.R.: Funciones cerebrales y psicopatología, México, D. F.: Alfíl, S. A. de C. V. (2013)
4. Rao, S.: Integration of what and where in the primate prefrontal cortex. *Science*, Vol. 276, No. 821 (1997)
5. Miller, G.: Reflecting on another's mind. *Science*, Vol. 308, No. 945 (2005)
6. Gaddes, W.H., Edgell, D.: *Learning Disabilities and Brain Function*. New York: Springer-Verlag, (1994)
7. Cohen, R.A., Sparling-Cohen, Y.A., O'Donnell, B.F.: *The Neuropsychology of Attention*. New York: Plenum Press (1993)
8. Ostrosky-Solís, F., Gómez Pérez, E., Chayo-Dichy, R., Flores Lázaro, J.C.: ¿Problemas de atención? Un programa para su estimulación y rehabilitación, D. F. American Book Store (2004)
9. Téllez Olvera, H., Mendoza González, M.E., Butcher López, E.A., Pacheco Ralley, C.C., Tirado Medina, H.: *Atención, aprendizaje y memoria. Aspectos psicobiológicos*. México, Trillas (2002)
10. Kirby, E., Grimley, L.: *Trastorno por déficit de atención*. México, Limusa-Noriega (1992)
11. Conners, C.K.: *Conners parents rating scale*. Chicago, Abbott Laboratories (1979)
12. Ardila, A., Ostrosky, F.: *Guía para el diagnóstico neuropsicológico*, D.F, México (2012)
13. Stern, W.: *Die Intelligenz der Kinder und Jugendlichen und die Methoden ihrer Untersuchung*. Leipzig Verlag (1928)
14. Ervin, F., Mark, V.: *Violence and the brain*. New York, Harper & Row (1970)
15. Monroe, R.: *Episodic behavioral disorders*. Cambridge: Harvard University Press (1970)
16. Corsi, P.M.: *Human memory and the medial temporal region of the brain*. APA PsycNet, Dissertation Abstracts International (1972)
17. Ostrosky-Solís, F., Gutiérrez Vivó, J.: ¡Toc, Toc!, ¿hay alguien ahí?. *Cerebro y conducta. Manual para usuarios inexpertos*. México: Info Red (2000)
18. Ostrosky-Solís, F., Ardila, A., Roselli, M., López, G., Mendoza, V.: Neuropsychological test Performance in Illiterates. *Clinical Neuropsychology* 13(7), 645–660 (1998)
19. Ostrosky-Solis, F., Ardila, A., Roselli, M.: *NEUROPSI. Evaluación Neuropsicológica Breve en español*. Publingenio
20. Ostrosky, F., Gómez, E., Matute, E., Roselli, M., Ardila, A., Pineda, D.: *Neuropsi, Atención y Memoria. Manual e instructivo*, México, American Book Store (2007)
21. Ostrosky-Solís, F., Gómez, E., Matute, E., Roselli, M., Ardila, A., Pineda, D.: *NeuroPsi, Atención y memoria. Manual, Protocolos, Láminas, Tablas Puntuaciones Totales y Perfiles*, México, American Book Store & Teleton (2003)
22. Wechsler, D.: *WMS-III Administration and Scoring Manual*. San Antonio, Texas, The Psychological Corporation (1997)

23. Lezak, M.D., Howieson, D. B., Loring, D. W.: Neuropsychological assessment. New York, Oxford University Press (2004)
24. Wechsler, D.: Wechsler Adult Intelligence Scale–Revised. New York, The Psychological Corporation (1981)
25. Gromwall, D.: Paced Auditory Serial Addition Task: A measure of recovery from concussion. *Perceptual and Motor Skills*, No. 44, pp. 367–373 (1977)
26. Reitan, R.M., Wolfson, D.: The Halstead-Reitan Neuropsychological Test Battery: Theory and interpretation. Tucson, Neuropsychology Press (1985)
27. Ardila, A., Rosselli, M.: Neuropsicología clínica. México, Manual Moderno (2007)
28. Ardila, A., Ostrosky-Solis, F., Rosselli, M., Gomez, C.: Age related cognitive decline during normal aging: The complex effect of education. *Archives of Clinical Neuropsychology*, No. 15, pp. 495–514 (2000)
29. Pixframe Studios: Towi: Desarrollo de habilidades del aprendizaje a través del juego, <http://towi.com.mx/> [Last access: 20 06 2017]
30. Martínez Jiménez, M.Á., Hernández Mendo, A., Pastrana Brincones, J.L.: Programa informático para evaluación y entrenamiento de la atención. <http://www.efdeportes.com/efd145/programa-informatico-para-evaluacion-de-la-atencion.htm> [Last Access: 20 06 2017]
31. Unobrain: Entrenamiento cerebral, <http://www.unobrain.com/entrenamiento-cerebral> [Last access: 20 06 2017]
32. CogniFit: Test Neuropsicológicos y Estimulación Cognitiva, <https://www.cognifit.com/es> [Last access: 20 06 2017]
33. Preiss, M., Shatil, E., Cermakova, R., Cimermannova, D.: Personalized cognitive training in unipolar and bipolar disorder: a study of cognitive functioning. *Frontiers in Human Neuroscience* (2013)
34. Haimov, I., Shatil, E.: Cognitive Training Improves Sleep Quality and Cognitive Function among Older Adults with Insomnia. *PLoS ONE* (2013)
35. Merete, Ø., Bjørn Rishovd, R.: Neuropsychological Deficits in Adolescent-Onset Schizophrenia Compared with Attention Deficit Hyperactivity Disorder. *Am J Psychiatry* 156(388) (1999)
36. Wilens, T.E., Spencer, T.J.: Understanding Attention-Deficit/Hyperactivity Disorder from Childhood to Adulthood. *JC Psychiat Suppl.* 67(7), 8 (2006)
37. Pérez-Montoro Gutiérrez, M.: Arquitectura de la información en entornos web. Asturias, España: Ediciones Trea (2010)
38. Ezequiel Rozic, S.: Bases de datos. Buenos Aires, Argentina, MP Ediciones (2004)
39. Oppel, A.J.: Databases: A beginner's guide. McGraw-Hill (2009)
40. Silberschatz, A., Korth, H.F., Sudarshan, S.: Fundamentos de bases de datos. Madrid, España, McGraw-Hill (2002)
41. Chodorow, K.: MongoDB. The definitive guide. California, USA, O'Reilly Media (2013)
42. Wilson, M.: Building Node Applications with MongoDB and Backbone. California, USA, O'Reilly Media (2013)
43. Pollice, G., Augustine, L., Lowe, C., Madhur, J.: Software development for small teams: RUP-Centric approach. Pearson Education (2004)
44. Leffingwell, D., Widrig, D.: Managing software requirements. A use case approach. Pearson Education (2003)
45. Jacobson, I., Booch, G., Rumbaugh, J.: El proceso unificado de desarrollo de software. Madrid, España, Pearson Education (2000)



# Body Sensor Network Using a Domotic System

Francisco Beltrán<sup>1</sup>, Felix Mata<sup>1</sup>, Mario Rivero<sup>2</sup>

<sup>1</sup> Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, SEPI-UPITA, Mexico City, Mexico  
beltranchavezjf@gmail.com, mmatar@ipn.mx

<sup>2</sup> Instituto Politécnico Nacional, Centro de Investigación en Computación, Laboratorio de Redes y Ciencia de Datos, Mexico City, Mexico  
erivero@cic.ipn.mx

**Abstract.** In this paper is presented the design of a domotic system with a body sensor network (BSN) based on an internet of things (IoT) architecture. It is presented two phases of the project: 1) a BSN evaluation based on computing simulation and 2) a system that collects the information generated by the BSN simulation and send to a Cloud system, to process the data and applying different analysis. One of the goals of this research is to define an IoT architecture that works as a framework for future implementations focused on sensor networks. Summing up, it is described a domotic system that monitors different physiological signals (e.g., cardiac rhythm) these data are sent to a cloud system. In the cloud data is processed and if an event in the data is detected (e.g. accelerated heart rate) then an action is triggered (e.g., turn on relaxing music). The BSN was simulated computationally. Simulation is validated and general outline of the system is presented and the future work is discussed.

**Keywords:** BSN, IoT, domotic, cloud environment.

## 1 Introduction

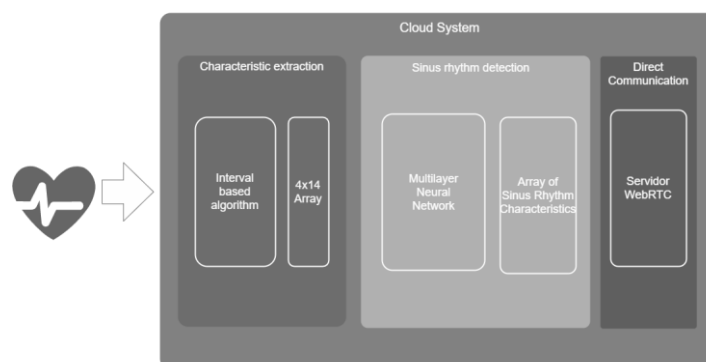
As stated by several sources, IoT has evolved into a state where it's not seen as an emerging technology by itself, but as a platform that evolves and can help other emerging technologies achieve their full potential [5]. However, the term itself involves many other technologies, most of them different from each other, and the definition can vary from author to author. As [4] mentions IoT has passed through three different stages, which can be characterized by the technologies involved in them. These stages mark different milestones that have improved the development of IoT systems, and that also have helped the definition of the term to become more concrete. However, the state-of-art systems that have been published in recent years don't seem to incorporate these three stages and in most instances they seem to incorporate tools that make impossible the interaction between them. Because of this, it's proposed the development of a simple IoT system that works as a guideline for future implementations that deal with similar technologies.

## 2 Methodology

The first two phases of this project are a simulated Body Sensor Network and a Cloud system to analyze the data generated by the BSN.

### 2.1 BSN Simulation

The simulation of BSN consists of the configuration required to many sensors acquire different data (physiological signs) from a person (or even many persons). The simulated BSN comprises a certain number of nodes that transmit physiological signs, extracted from a person. The communication and transferring data is through the use of the protocol Slotted ALOHA for medium access control. The binary data transmission is used to apply pattern recognition (identify an event into the signal, for example: accelerated heart rate). Fig. 1 shows a general outline of the system's structure.



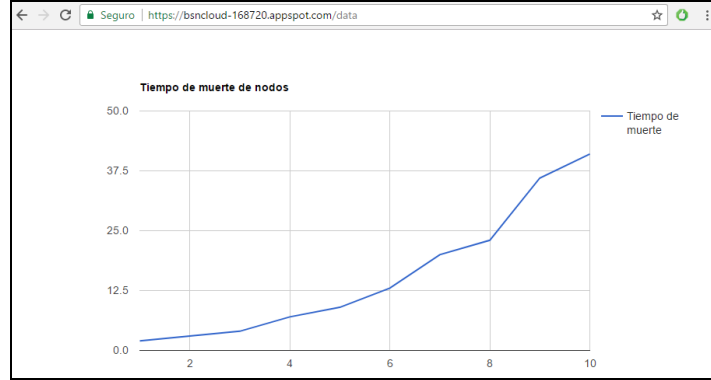
**Fig 1.** Outline of the system's structure. As can be seen, the cloud system will be comprised of three different modules.

### 2.2 Cloud System Communication with Energy Saving

The Cloud system receives the data of vital signals generated by the BSN simulation. These data are analyzed to identify or detect an event (e.g. characteristic extraction). The data is received through a gateway device that coordinates the data flow from the BSN to the Cloud. Here, one of the goals is save energy in the communication process. To achieve that, the BSN coordination is assigned to a device with higher capabilities within the BSN nodes. It ensured that the nodes save energy by only using it to communicate with the data sink and their own functionality needs.

The cloud system is envisioned as a series of web services that allow any kind of device that can implement the http protocol to make use of them through the use of an URL. In the tests done the cloud environment server is comprised by four web services that allow the BSN to send data to the Neural Network located on the server.

Fig. 2 shows a graph that displays the death times of the nodes in the BSN. Said test displays the result of the calls made to a web service by the simulated BSN.



**Fig 2.** Transmission test from the BSN to the cloud environment with a graph showing death time for the net nodes.

### 2.3 Communication in BSN Simulation

The BSN was simulated as a set of nodes that transmit information to a sink under a Slotted Aloha protocol. Using a simulation proved to be easier to implement than a real testbed with nodes attached to real people since in order to determine the different operational parameters of the BSN or the nature of the obtained signal we can just simply change a few lines of code. The physiological signals used in this project were taken from The MIT-BIH Normal Sinus Rhythm Database which contains physiological signals from healthy and abnormal sinus rhythms [7]. In Fig. 3 the workflow of the simulation is shown.

The simulation's performance was validated using a Markov chain denoted by equation 1 and a comparison of the throughput obtained through the simulation and the Markov chain is shown on Fig. 4 and 5.

$$S = \sum_{i=0}^N \frac{1}{i\tau(1-\tau)^{i-1}} = \frac{1}{\tau} \sum_{i=0}^N \frac{1}{i(1-\tau)^{i-1}}, \quad (1)$$

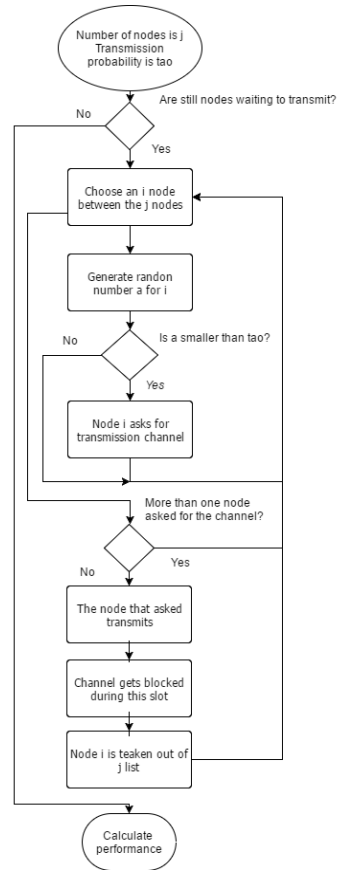
where:

$S$  = throughput

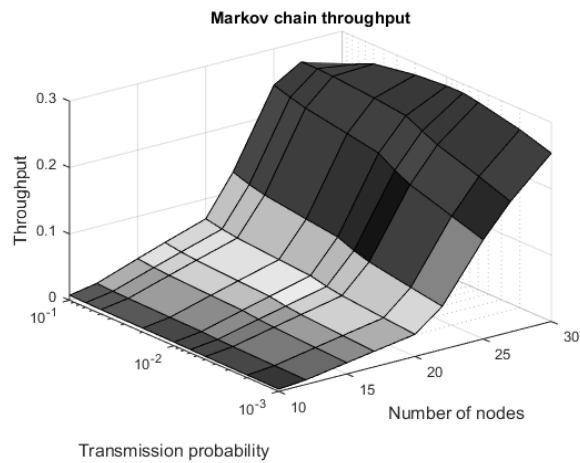
$N$  = number of nodes

$\tau$  = transmission probability

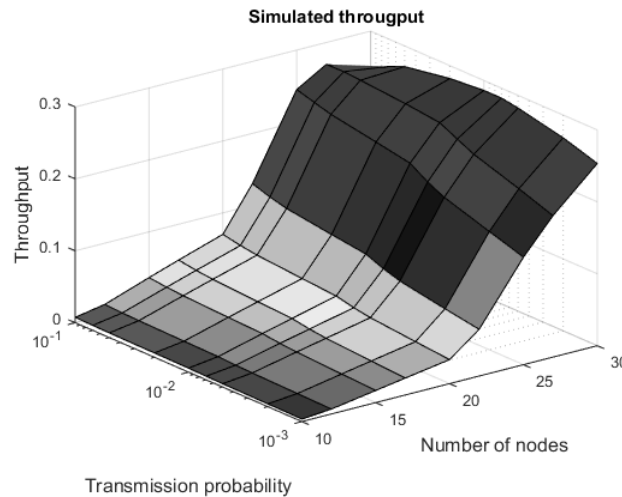
It is shown that the throughput reaches its maximum value when the transmission probability is selected around  $1 \times 10^{-2}$  and it is proportional to the number of nodes in the BSN. It has also been proposed an energy aware transmission protocol. This protocol assigns each node a different transmission probability based on its energy left and based on the networks' residual energy, however, testing has shown that this leads to nodes not transmitting at all when energy levels are low.



**Fig. 3.** Workflow of the BSN simulation.



**Fig. 4.** Throughput obtained by solving Equation 1.



**Fig. 5.** Throughput obtained by the simulation.

The cloud system as for this stage, is comprised of four different Restful web services created with the Java library Jersey. These web services correspond the CRUD paradigm and serve as the only means to communicate with the server. They were accessed from the BSN through the use of the *curl* tool.

The characteristic extraction is based on [2], where the classification of the electrocardiogram signal is based on 13 different characteristics extracted from each heartbeat and the relation between heartbeats. The extracted characteristics can be classified under three categories: qrs complex characteristics, inter beat relations and time related characteristics. These characteristics are given as a  $1 \times 14$  array input to a two layered neural network. Work is still being done on this module since its implementation relies heavily on WFDB Toolbox on Matlab [6,7].

### 3 Experiments

The test was an implementation of the character recognition exercise given by [3] for testing the connection between the BSN and the server. In said test an array of 37 different values are given to a neural network in order to be classified according to the letter they represent. Through the access to an URL, the BSN uploaded bipolar arrays that represent characters and the server classified them based on their similitude to the correct characters.

The web services are four, but for this test only one was used. The test was successful and given that the sole purpose of the test was to ensure the correct connectivity between the BSN and the server it is concluded that the system passed the test.

## 4 Future Work

In the first place, it has been stated that the pattern recognition method used for the tests so far, although it has proven to be successful, it is still not validated as the best method for this kind of implementations. Since detecting fluctuations on the normal sinus rhythm, is the main focus of this project more work is needed in order to get a more refined method of detection. Further testing is needed in order to get a higher percentage of detection.

Secondly, the tests done so far, involve only one simulated BSN. Further tests must involve more simulated BSNs in order to test the Cloud environment stability under a big number of BSNs sending data continuously. Also, the patterns transmitted from the BSN to the server must be physiological signals in future tests instead of bipolar arrays.

Finally, the BSN simulation is focused mainly on the transmission behavior of the net. In future tests the BSN must be modeled as close as possible to a real BSN.

## References

1. Atzori, L., Iera, A., Morabito, G.: Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Networks* 56, pp. 122–140 (2017)
2. Darouei, A., Ayatollahi, A.: Classification of Cardiac Arrhythmias Using PNN Neural Network Based on ECG Feature Extraction. In: *World Congress on Medical Physics and Biomedical Engineering*, pp. 309–311 (2009)
3. Fausett, L.: *Fundamentals of neural networks*. Pearson, pp. 59–66 (1994)
4. Fortino, G., Di Fatta, G., Pathan, M., Vasilakos, A.: Cloud-assisted body area networks: state-of-the-art and future challenges. *Wireless Networks* 20, pp. 1925–1938 (2014)
5. Reinfurt, L., Breitenbücher, U., Falkenthal, M., Leymann, F., Riegg, A.: Internet of things patterns. In: *Proceedings of the 21<sup>st</sup> European Conference on Pattern Languages of Programs* (2016)
6. Silva, I., Moody, G.: An Open-source Toolbox for Analyzing and Processing PhysioNet Databases in MATLAB and Octave. *Journal of Open Research Software* 2(1), e27 (2014)
7. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101(23), E215-20 (2000)

# Price Forecasting: A Data Science Approach

Paul Ricardo Millán González, Félix Mata

Instituto Politécnico Nacional, UPIITA, SEPI,  
Mexico City, Mexico

luapmg@outlook.es, mmatar@ipn.mx

**Abstract.** The research presents a data science approach to analyze prices of products in supermarkets to make price forecasting and discover patterns of behavior in the prices. The dataset comprises a period from 2011 to 2014. The case of study selected in this research is, a price forecasting experiment analyzing data of tuna product in Mexican republic. Data were processed and classified using the methods of linear regression, polynomial regression, regression support vector and neural networks. This research provides preliminary results as a first advance of the approach.

**Keywords:** Machine learning, data science, prediction.

## 1 Introduction

Data Analysis is used in this research as an approach to work with inference to derive the conclusion based on previous experience of data behavior. It means find or define mathematical models and algorithms from the data.

The family economy is very important, hence is important to be informed about the behavior prices of basic products principally, such as food, home cleaning products and health care products. Currently exist a public database offered by PROFECO [3] with more than 27 million of records of different products offered in distinct establishments and commercial chains. This database includes attributes like product, presentation, brand, state, municipality, address, date and price.

This research will focus on to analyze the behavior of price for product tuna in Mexican Republic. Our hypothesis is based on the assumption that is possible to create a mathematical model that describe the price behavior of it over the time, it will be achieved using techniques of Machine Learning. This mathematical model can be useful to forecast or discover patterns regarding to price behavior for certain products in the basket basic.

As a first step, it was considered that in a small data sample will be enough to find a model. It was based in a previous analysis when the data was inspected followed the approach of data mining, using tools like RapidMiner (<https://rapidminer.com/>). It was found, for example, some cases where two supermarkets of the same branch and even in the same city and at the same time presented different prices for a particular product. Therefore, it was thought that every establishment could had your own behavior, and this would add complexity to the experiment because although the database contains

thousands of records these are distributed around all Mexico's territory. Nevertheless, when data were fragmented by city or state the size is decreased, then it was not possible to use them for finding a model.

The rest of the paper is organized as follows: Section 2 describe the related work, in section 3 the methodology used is explained, in section 4 the experiments are done, in section 5 the results are shown and section 6 outlines the conclusions.

## **2 Related Work**

There is a large amount of forms in forecasting economy indices such as stock prices [1], in which the methods used are time series analysis, and the majority are focused on structured data (e.g. stock prices table) [2,3].

Many researchers are doing about the use of machine learning for forecasting, many areas of study can be benefited with this approach. In business it is possible to predict with an acceptable range of error the price behavior of one particular product. Several works use approaches based on machine learning, concretely, it is a matter of creating programs capable of generalizing behaviors from information provided in the form of examples. It is, therefore, a process of induction of knowledge [1].

Other related work is [4] focused on predict the price of natural gas, but not using a data science approach like in our work. While, other works like [5] where is compared support vector machines for regression with Back propagation and RBF networks in stock price data. Finally, the main motivation of our research is the possibility to help, for consumers, companies and government, in the creation of strategies of purchase or public policies.

## **3 Methodology**

The methodology defined consist of three experiments: 1) A Small DataSet (only canned tuna), 2) A Big DataSet with tuna per kilogram and 3) A Big DataSet (only canned tuna).

In the case of experiment 1) with small dataset. The sample was 101 records of the branch Tuny, 140grams in oil, for store Soriana Branch, Mexico City, in Azcapotzalco from 2011 to 2013. The only components in this stage of the study were the date and the price, the date attribute was treated in different ways, first it was transformed the date that came in chain format in three integer values day, month and year, after being transformed each one of these variables in ranges, then the whole date was transformed into rank, finally, the date was transformed into a special continuous float value for regression analysis. The different ways in which the date attribute was transformed did not generate variation in the effectiveness of the model therefore we decided to use the date as a continuous float value as it gives us the advantage that we can print the data easily.

For the case of experiment 2) and 3) big dataset. The sample was 169469 records; this sample data was used to find an acceptable model and a good performance when machine learning algorithms were applied. Data was raised from distinct states and cities around Mexico. The classification consisted of groups such as: state, municipality

and commercial chain. Combinations were performed only with those have more than 100 records. There are two Big datasets, one with canned tuna and tuna per kilograms then this dataset has low prices (canned tuna) and large prices (tuna per kilogram), and another with only canned tuna is that the reason that this only has lower prices.

## 4 Machine Learning

### 4.1 Linear Regression with Small DataSet

It is beginning by training a linear regression model: The method used for regression is expected to be a linear combination of the input variables.

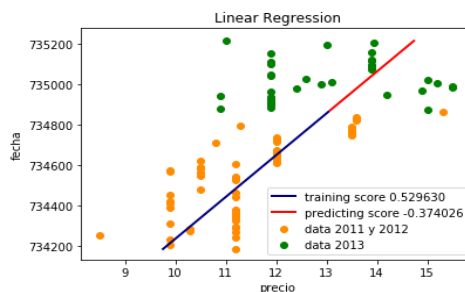


Fig.1. Lineal Regression.

In Fig. 1 it is possible to see a model greeted by lineal regression nevertheless this model does not represent well the behavior of the sample.

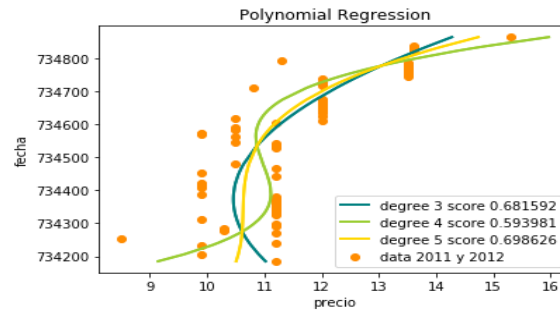
### 4.2 Polynomial Regression with Small DataSet

A new feature matrix is generated consisting of all polynomial combinations of entities with a degree less than or equal to the specified degree.

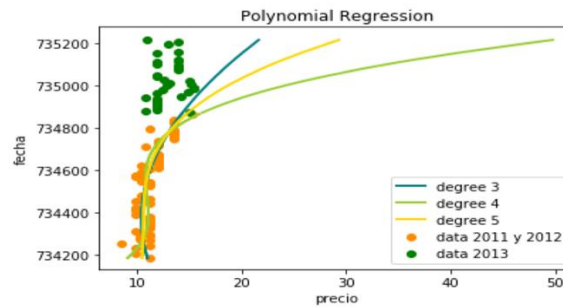
In Fig. 2 the models represent well the sample, but it is only at the training. While in Fig. 3 can be appreciated that these models do not predict correctly.

### 4.3 SVR with Small DataSet

Given a set of points, a subset of a larger set (space), each of which belongs to one of two possible categories, an algorithm based on SVM constructs a model capable of predicting whether a new point (whose category we do not know) belongs to one category or the other. The SVM looks for a hyper plane that optimally separates the points of a class from the one of another, that could have been previously projected to a space of superior dimensionality. The support vector machines for regression is a robust technique for function approximation. [5].



**Fig. 2.** Polynomial Regression Training.



**Fig. 3.** Polynomial Regression Predicted.

At the moment the algorithm SVR is the one that appears to have the best behavior, but it is not enough. Several tests were done modifying the values of gamma, constant C and epsilon. Fig. 4 shows one of the best results, which as we can see has a negative sign which is very bad since this means that the model does not represent the behavior of the data. From this moment it was considered that the sample was too small and it was proposed to make tests with more data. The Fig. 4 show the results obtained by linear regression and SVR.

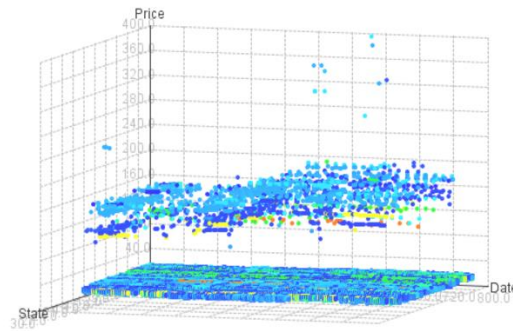


**Fig. 4.** Price forecasting over time.

#### 4.4 Neural Networks with Big DataSet

Broadly speaking, two views exist between practitioners/investors who typically prefer a small in-sample to minimize data holding requirements and researchers/academics who typically chose large in-sample periods [6].

In this stage, the data sample contains 169469 records, it was grouped by state, municipality and commercial chain and took only those that combinations that had more than 100 records. This in reason that was considered that those combinations that had less of 100 could decrease the performance of training as there were not enough samples. The experiment was performed using Microsoft Azure. In Fig. 5 it can be seen the behavior of the Big DataSet with tuna per kilogram over the time. When we talk about the complete sample we refer to the first sample that was selected, which included tuna per kg. Thus, we have low prices of canned tuna and high prices of tuna per kilogram.



**Fig. 5.** Big DataSet with tuna per kilogram.



**Fig. 6.** Big DataSet (only canned tuna).

It was used neural networks because at the beginning it was observed that this algorithm presented the best results. Therefore, the research is centered hereafter applying neural networks. In this sample we have the following attributes: presentation, brand, commercial chain, state, municipality, address all these are string. Date, price: double.

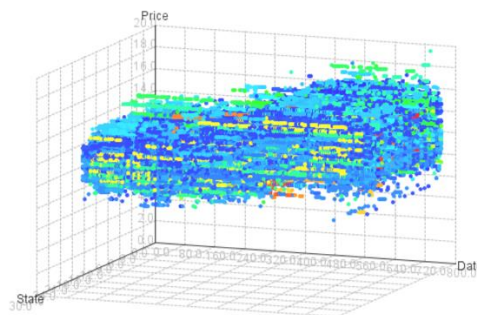
From the previous attributes new attributes were made as day: integer, month: integer, year: integer, day\_week: string, month\_chain: string. It was tried to use one hot code to binarize the text columns presentation, state, municipality, address, however many columns were created in the process, which complicated the treatment of the sample, therefore this option was declined.

The parameters of Neural Network Regression module were those of default. Create trainer mode: single parameter, Hidden layer specification: fully-connected Number of

hidden nodes: 100, Learning rate: 0.005, Number of learning iterations: 100, The initial learning iterations: 100, The initial learning weights diameter: 0.1, the momentum: 0, The type of normalizer: Min-Max normalizer.

Now there are obvious questions like what would happen if large prices (prices of tuna per kilogram) were removed? And is valid keep them? are these samples helping at the training really? or it is just a coincidence and these samples only confuse the score of performance in the training stage?

Now, large prices (prices of tuna per kilogram) will be removed, and the training will be done one more time, the same columns will be used and the same parameters of the algorithm. In the Fig. 7 is the price behavior over the time of the new sample without large prices. In this figure it can be seen a tridimensional plot of this sample on time, state and price, the colors represent chains of store.



**Fig. 7.** Big DataSet (only canned tuna) on Date, State and Price.

## 5 Results

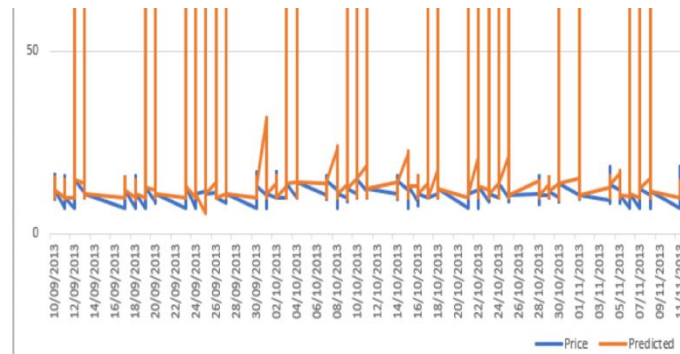
### 5.1 Results with Big DataSet with Tuna per Kilograms

These are the results of the sample with large prices (with tuna per kilograms), the output provided by Azure with Big DataSet with tuna per kilograms are the followings: Mean Absolute Error = 0.996136, Root Mean Squared Error = 1.727647, Relative Absolute Error = 0.375668, Relative Squared Error = 0.019779, Coefficient of Determination = 0.980221. The Mean Absolute Percentage Error (MAPE) is 7.75%. The maximum error is 201.27% The number of predictions with error above 20% is 1047 of 16947 so 6.17% is outside a range of error that can be considered acceptable.



**Fig. 8** Price Behavior Over the Time for Big DataSet with tuna per kilogram.

The Fig. 9 above show the real price behavior over the time against the model provided from our methodology using Neural Network Regression. To help the reader to see better, a zoom is shown in the Figs 10 and 11.



**Fig. 9** Zoom of graph in Fig. 8.

The Fig. 9 above is the first part of the zoom of the Fig. 9, here is easier to appreciate the difference between the real price and the predicted price.

## 5.2 Results with Big DataSet (only canned tuna)

This is the sample without large pri, the result when the large prices of tuna per kilogram are put away is that the performance decline completely. The results provided by Azure for Big DataSet (only canned tuna) are the followings: Mean Absolute Error = 1.000164, Root Mean Squared Error = 1.394822, Relative Absolute Error = 0.707804, Relative Squared Error = 0.71398, Coefficient of Determination = 0.28602. The mean absolute percentage error (MAPE) is 11.52%. The maximum error is 83.01%. The number predictions with error above than 20% are 3023 of 16699 so 18.10% is outside a range of error that can be considered acceptable, which is threefold of the previous result.



**Fig. 10.** Price Behavior Over the Time of Big DataSet (only canned tuna).

## 6 Conclusions

Is interesting to see that the large prices of tuna per kilograms can help to obtain better results on prediction, though this is on discussion because it produced bigger errors too.

In my opinion if it is possible to improve the result of predictions using cross data, it would be valid to use them. Many relations can exist between our data and other information available outside. In this case is obviously that exist a relation between tuna and canned tuna, but the relation seems be linear, so is interesting to see how much seem to help at the training more data that seem to behave same, but in distinct range of prices.

More tests are necessary to be able to lower the mean percentage error(MPE) and thus make better predictions. We believe that the model can be improved in such a way that the predicted behavior will be closer to real, there is the possibility that it could be achieved with more samples if we could get an upgrade of our dataset that contain more years or by adding data from other sources of information.

Both tests in the second stage offer good results, that one is better than another depends on the context where you want to apply.

The opinion of the author of this paper is that the experiment of all presentations of tuna with large and low prices together yield better results, because although it has bigger mistakes these are few and the most of predictions are close to real price. However, it is a preliminary study and therefore this is still a supposition, a much larger study is necessary to provide better results and to be able to sustain them.

## References

1. Stock, J., Watson, M.: Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, vol. 41, pp. 788–829 (2001)
2. Marcek, D.: Stock price forecasting: Statistical, classical and fuzzy neural network approach. In: V. Torra and Y. Narukawa (eds.), *MDAI*, ser. *Lecture Notes in Computer Science*, vol. 3131. Springer, pp. 41–48 (2004)
3. A hybrid ARIMA and support vector machines model in stock price forecasting, vol. 33, no. 3, (2005)
4. Dataset source: <https://datos.gob.mx/busca/dataset/quien-es-quien-en-los-precios/resource/9fa38cc3-bcc6-4597-b240-263532532467> [Last access: 10 2017].
5. Čeperić, E., Žiković, S., Čeperić, V.: Short-Term Forecasting of Natural Gas Prices using Machine Learning and Feature Selection Algorithms. *Energy* (2017)
6. Trafalis, T.B., Ince, H.: Support Vector Machine for Regression And Applications to Financial Forecasting. In: *IJCNN '00 Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)*, volume 6 (2000)
7. Kambouroudis, D.S., McMillan, D.G.: Is there an ideal in sample length for forecasting volatility? *Journal of International Financial Markets, Institutions and Money* (2015)

Electronic edition  
Available online: <http://www.rcs.cic.ipn.mx>



ISSN: 1870-4069  
<http://rscs.cic.ipn.mx>



Centro de Investigación  
en Computación