EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 151 No. 12
December 2022

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:
Griselda Franco Sánchez

Research in Computing Science, Año 21, Volumen 151, No. 12, diciembre de 2022, 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 diciembre de 2022.

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 21, Volume 151, No. 12, December 2022, 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.

Volume 151(12)

Advances in Artificial Intelligence and Industry 4.0

Edgar Gonzalo Cossio Franco
Humberto Sossa Azuela
Gustavo Trinidad Rubín Linares (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2022

ISSN: in process

Copyright © Instituto Politécnico Nacional 2022
Formerly ISSNs: 1870-4069, 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

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Evolutionary Tuning of a Fuzzy Controller for Electrolyte Flow Regulation in a Pulsed Electrochemical Machining Process

Irvin Uriel Nopalera-Angeles, Everardo Efrén Granda-Gutiérrez, René Arnulfo García-Hernández, Ángel Hernández-Castañeda, Juan Fernando García-Mejía

Universidad Autónoma del Estado de México,
Instituto Literario,
Mexico

irvin_12f@hotmail.com, {eegrandag, reagarciah, anhernandezc, fgarciam}@uaemex.mx

Abstract. Based on Darwin's theory of natural selection and the laws of inheritance proposed by Mendel, a stochastic optimization established by a real-coded genetic algorithm is described for tuning the parameters of the membership functions of a multiple-input, single-output Mamdani-type fuzzy controller. This control is applied in modern manufacturing, which regulates the electrolyte flow in a pulsed electrochemical micromachining method. In this optimization technique, the results generated with a population crossover of 60 and 80 %, using a roulette selection, a BLX-α crossover operator, and a uniform mutation are analyzed. In addition, the electrolyte flow control response of the best-fit chromosomes according to an objective function set by the mean square error is compared to the output obtained with a PID controller tuned with Ziegler Nichols. Finally, the control space for the fuzzy controller is generated with the parameters of the membership functions that offer the best performance in flow stabilization.

Keywords: Genetic algorithm, fuzzy control, electrolyte flow, modern manufacturing.

1 Introduction

Pulsed electrochemical machining (PECM) is derived from the non-conventional manufacturing process, which uses the principle of electrolysis to wear high-strength metal parts. In general terms, the dissolution process is catalyzed by the polarization of electric energy positively at the anode and negatively at the cathode when exposed to the electrolyte flow, causing the transfer of ions from the workpiece to the tool; both are set at a constant separation.

Therefore, a pair of electrodes (anode and cathode), a mobilization system for the working tools, an aqueous solution (electrolyte), an electrolyte flow system, and a pulsed polarization source are the elementary components that allow controlled

regulation of material wear [1]. Minimal tool wear, material removal on high-strength metal parts, and increased efficiency in manufacturing components with complex morphologies are part of the advantages offered by PECM.

Some of the applications of this technique are visualized in the aeronautics field for manufacturing LPC (Low-Pressure Compressor) blades, combustion chambers, diffusers, and cooling film. However, it is also used to design microtubes, micro gears, micro bushings, and scalpels [2-3].

Since the contact between the tool and the workpiece never exists in this manufacturing method, the electrolyte flow plays an important role in ensuring material wear. However, because of the detachment of metal, particles are generated that can cause an obstruction in the circulation of the solution, decreasing the flow and consequently the current transfer, thus affecting, mainly, the uniform detachment of material and the accurate estimation of the control variables [1].

Problems such as alterations in the final dimensioning of the finished part, increased electrical resistivity properties, wear of the working tool, and decreased technological stability are some of the effects identified by the deficient transfer of the solution [4].

Therefore, the correct control of the electrolyte supply is considered an essential factor in this process, which can be done through classical control methods or intelligent techniques with industrial applications such as fuzzy logic.

Fuzzy logic is a method derived from soft computing with applications in the design of industrial process control, which allows the evaluation of actual generic parameters using linguistic variables [5]. These linguistic labels are defined through a series of sets established in a universe of discourse representative of a solution space.

Therefore, being a knowledge-based system, the quality of the algorithm results will depend on the expertise expressed utilizing a knowledge base and the membership functions [6]. Additionally, the parameters that make up this type of intelligent control are classified into two groups: structural and tuning parameters [7].

The first group includes the input and output variables, the inference system, the rule base, and the defuzzification method. In contrast, the second group comprises the parameters of the membership functions, which are precisely delimited empirically or by search methods.

Genetic Algorithms are stochastic techniques helpful in searching for solutions based on evolutionary mechanisms from Darwin's theory of natural selection and the laws of inheritance proposed by Mendel [8], applicable in specialized problems from a random set of solutions called population. Some chromosomes delimit a particular solution to the problem in each population element.

After a series of iterations, commonly called generations, the algorithm evaluates the quality of the results with an objective function to subsequently select, cross, mutate and reconstruct the population with the best set of solutions found as well as with the descendants generated by them [9].

Hybridization of intelligent techniques is a way to improve the results from individually applied methods. This is observed in different fields of specialized literature as described in [10] and [11], where genetic algorithms are used to tune the parameters of the membership functions of fuzzy controllers.

In the robotics area, as observed in [12, 13] the tuning of the Kp, Ki, and Kd coefficients of classical controllers such as Proportional Integral Derivative (PID) but

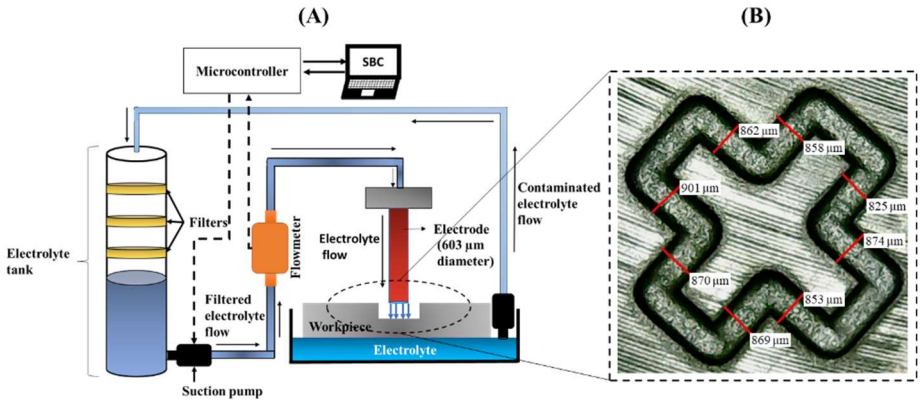


Fig. 1. General diagram of the electrolyte flow regulation system in the PECM manufacturing process (A) for micrometer-scale machining design (B).

also of fuzzy PIDs controllers is performed, validating the performance provided by the stochastic search of evolutionary algorithms.

In this sense, this paper describes the development of a genetic algorithm to tune the parameters of the membership functions of a fuzzy controller which regulates the electrolyte flow in a pulsed electrochemical machining process. This comprises two input parameters and one output, delimited by three membership functions of type Z, S, and Gaussian for each linguistic variable.

2 Methodology

The proposed evolutionary tuning is applied to an electrolyte flow regulation system implemented in a pulse electrochemical machining prototype described in [14], located in an experimental test laboratory.

It is composed of a hollow steel electrode with a diameter of 603 µm, a solution circulation system with suction pumps operating at 12 V - 1 A, and an electrolyte of Na at a molar mass concentration of 16 % per liter of H_2O. In Fig 1, the diagram of the components that make up the electrolyte circulation system in the non-conventional manufacturing process is shown.

An open-loop characterization was performed to determine the natural electrolyte flow rate Y during material removal. This relationship is described by a characteristic notation defined in equation 1, which represents the magnitude of the flow as a function of an oscillation frequency f caused by the passage of the solution in the internal mechanism of the flowmeter and a conversion factor k bounded by the volume of liquid supplied during a period:

$$Y = \frac{f}{k}. \tag{1}$$

A total of 1.182 liters of electrolyte in 60 minutes were generated in the characterization. Therefore, clearing k in equation 1 gives a conversion factor equal to

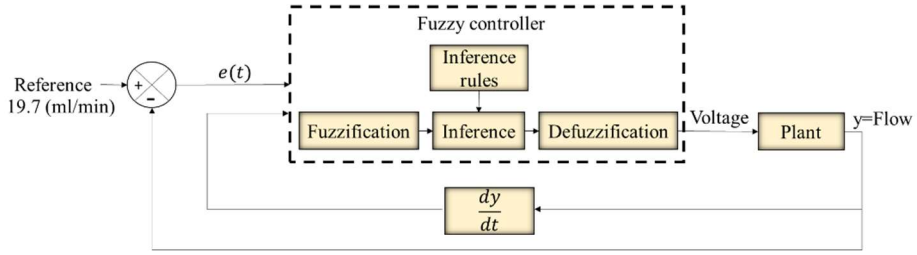


Fig. 2. Fuzzy control diagram for electrolyte control.

Table 1. Fuzzy associative memory for electrolyte control in PECM.

Fuzzy input sets	Minimum	Half	Advanced
Low	Little	Little	Little /Regular
Middle	Regular	Regular	Regular
High	Much/regular	Much/regular	Much

0.964 Hz*min/ml. Finally, substituting the values obtained in equation 1, a magnitude of 19.7 ml/min is generated in the solution flow rate during material removal. This behavior was analyzed using the MATLAB system identification module, considering the flow quantified during the characterization stage and the operating voltage of the suction pumps, establishing the third order transfer function in terms of the complex variable *s* of equation 2, which will allow modeling the system for its implementation in the evolutionary algorithm:

$$G(s) = \frac{1.203\,s\ + 3.099}{s^2\ +\ 1.534\,s + 0.945}. \tag{2}$$

The proposed evolutionary tuning is applied to a two-input, one-output fuzzy controller. The electrolyte flow error and the change of flow with respect to the time are the inputs evaluated in a fuzzy system composed of three stages: fuzzification, inference, and defuzzification. In the first one, a degree of membership of the actual variables to the fuzzy input is obtained.

In the second one, an inference is generated based on a series of IF-THEN type rules and the assigned sets for the output variable. Finally, in defuzzification, the fuzzy conclusion is converted to a numerical value interpreted by the non-conventional manufacturing system called "plant" according to the terminology in control. This allows the controller to modify its behavior through a closed-loop feedback system, as shown in Fig. 2.

Low, Middle, and High are the linguistic variables used to evaluate the first input variable defined by error. At the same time, Minimum, Half, and Advanced are the terms assigned to assess the change in flow with respect to time. On the other hand, the labels designated for the inference output are Little, Regular, and Much, in charge of delimiting the magnitude of the voltage applied in the plant.

In all cases, the terms are represented by Z, S, and Gaussian functions, respectively, using the fuzzy associative memory presented in Table 1. This describes the membership relationship of the fuzzy input sets located in the main column and row of Table 1 and the fuzzy output sets located in the central part of it.

Table 2. Membership functions for input variables.

Variable	**Fuzzy sets**		**Membership function parameters**
er (error)	LO	Low	μ-z (er; e_1, e_2)
	MI	Middle	μ-g (er; e_3, e_4)
	HI	High	μ-s (er; e_5, e_6)
dy (flow change)	MN	Minimum	μ-z (dy; e_7, e_8)
	HA	Half	μ-g (dy; e_9, e_{10})
	AD	Advanced	μ-s (dy; e_{11}, e_{12})

Table 3. Membership functions for the output variable.

Variable	**Fuzzy** sets		**Membership function parameters**
vo (voltage)	LI	Little	μ-z (vo; e_{13}, e_{14})
	RE	Regular	μ-g (vo; e_{15}, e_{16})
	MU	Much	μ-s (vo; e_{17}, e_{18})

Any function is composed of two indispensable parameters: the point of descent from 1 and the intersection at 0 for Z-type functions, the rising point from 0 and the intersection at 1 for the S-type, and finally, the standard deviation and the center for the Gaussian function. Both cases are shown in Table 2 for the input variables and in Table 3 for the output variable, set randomly initially and optimized by the evolutionary algorithm.

A structure of 18 genes for each chromosome in the population of the evolutionary algorithm is proposed based on the above. In this sense, each pair of genes will represent the parameters of the membership functions to be optimized, keeping the form of equation 3:

$$\text{Chromosome} = [e_1, e_2, e_3, e_4, \dots, e_{18}]. \tag{3}$$

In addition, an initial population with 100 randomly generated chromosomes was established, maintaining a universe of discourse from -25 to 25 ml/min for the error genes. On the other hand, gene initialization for flux change were trained under the same method with a universe of discourse from -30 to 30 ml/min^2. Finally, a population was randomly set from 0 to 12 V for the elements of the output variable. Once the population was generated, an objective function was established considering the mean square error of the electrolyte flow described in equation 4:

$$f_{obj} = \max\left(\frac{1}{1 + \sqrt{\frac{1}{T}\int_0^T (reference - output)^2}}\right). \tag{4}$$

The roulette operator is proposed for the selection of the chromosomes with the best fitness. At the same time, the generation of offspring is performed with a BLX-α cross, which allows creation a random offspring from the combination of the genes of two-parent chromosomes and a uniform parameter between 0 and 1 called alpha, as shown in equation 5:

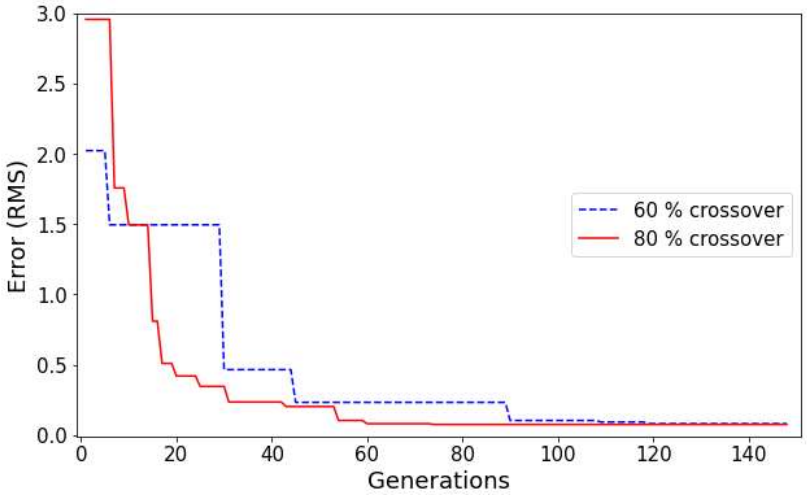


Fig. 3. Convergence of the algorithm with 60 and 80 % crossover of the population.

Table 4. Statistical analysis of the genetic algorithm with the 60 and 80 % crossover treatments in the population.

Technique	Crossover %	Error average	Standard deviation	K-S test (P-value)
GA	60	0.0825	0.008156	.09395
GA	80	0.0743	0.010847	.37194

$$D_n = \text{random}[(G_{min} - R * \alpha), (G_{max} + R * \alpha)], \quad (5)$$

where D_n is the chromosome generated as offspring, G_{min} is the minimum gene value of the parents $[G^1, G^2]$, G_{max} is the maximum gene value of the parents $[G^1, G^2]$, R is the difference of $G_{max} - G_{min}$ and α a random value between [0-1]. Finally, a uniform 5% mutation is applied on a randomly selected population during each generation of the algorithm, applying a comparative study between the results generated with a 60 and 80 % cross, verifying their distribution with the Kolmogorov-Smirnov test and, if necessary, applying the ANOVA test to demonstrate if they present significant differences between the two treatments. This is also contrasted with the results obtained with a PID controller tuned with the Ziegler Nichols method, establishing the value of the parameters $K_p = 0.61895$, $K_i = 0.60241$ and $K_d = -0.0036784$.

3 Results

This section shows the results obtained from the search for the parameters of the membership functions using the genetic algorithm. First, Fig. 3 shows the convergences obtained from the crossover with 60 and 80 % of the population in a total of 150 generations, showing that the evolutionary algorithm converges earlier in generation 76

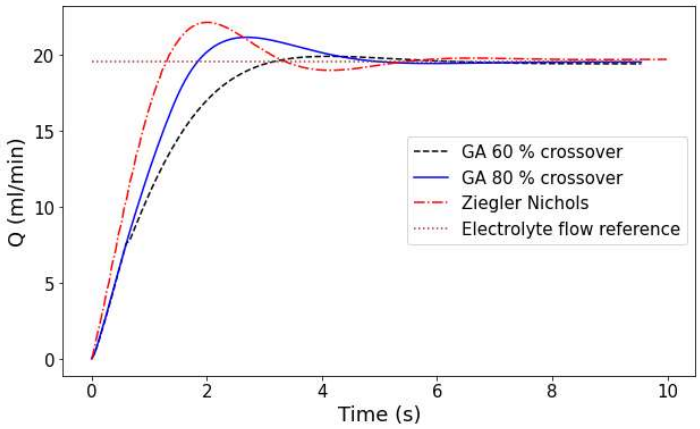


Fig. 4. Stabilization of electrolyte flow with the chromosomes with the best fitness in the search with 60 and 80 % crossover of the population, in addition to the response generated by the PID controller tuned with Ziegler Nichols.

Table 5. Performance criteria of the tuned controller with a genetic algorithm and Ziegler Nichols.

Test	Technique	% Crossover	RMS error	% Overshoot	Stablishment time (s)	Stable state error
1	GA	60	0.07988	1.50	5.8	0.14118
2	GA	80	0.07448	7.88	4.6	0.11997
3	PID (Ziegler)	-	0.10465	12.87	7.2	0.42549

for the criterion with the highest crossover percentage with a mean square error of 0.07448, while the convergence obtained for 60 % of the population was established in generation 121 with an RMS error of 0.07988.

As a result of the above, a statistical analysis was performed to determine the characteristics of the genetic algorithm when implementing both treatments in the proposed optimization. The results obtained are described in Table 4.

As shown in Table 4, the result of the Kolmogorov-Smirnov analysis indicates a normal distribution behavior for both experiments because the significance value is greater than 0.05.

Therefore, after applying the ANOVA test, a value of 0.006 is obtained, establishing that there are statistically significant differences between the treatment with different percentages of crossover between populations.

Subsequently, the chromosomes with the best aptitude from the previous experiments were extracted to be applied to the plant, and the response in the electrolyte flow control was analyzed. Furthermore, these results are contrasted with the stabilization of the controller tuned with the Ziegler Nichols method.

This is observed in Fig. 4, which describes each case for regulating the flow rate to a reference 19.7 ml/min. In addition, Table 5 describes the evaluation criteria defined

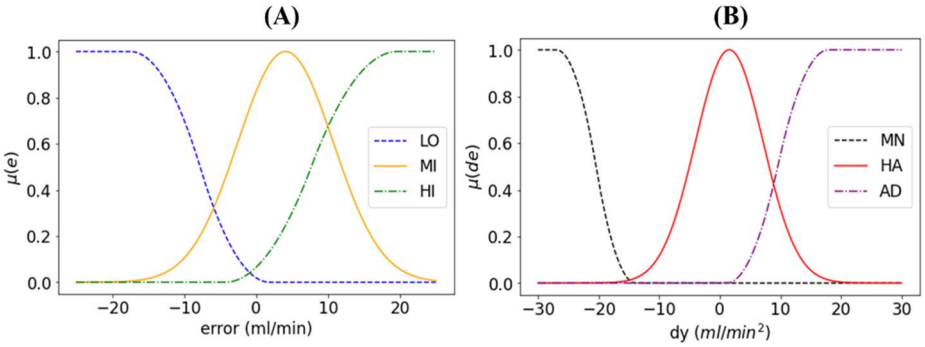


Fig. 5. The membership functions generated with the chromosome with the best fitness for evaluation of (A) error and (B) flux change with respect to time.

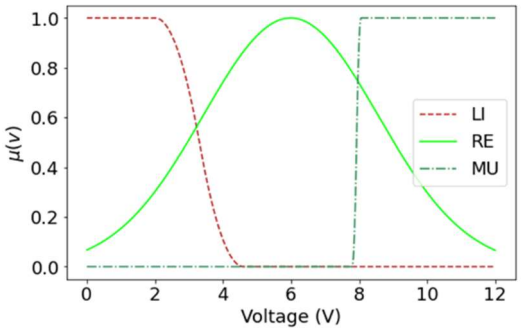


Fig. 6. Membership functions for the output variable.

according to the percentage of overshoot, steady-state error, and settling time for every tuning method.

As shown in Table 5, the genetic algorithm with a crossover percentage of 80 has a lower steady-state error and RMS than in the other cases. It requires less time to reach the flow establishment in error +/- 2 %.

Although the genetic algorithm of test 1 presents less over impulse, it is considered less significant than the response of test 2 since the establishment time with the lowest error is prioritized. In this sense, the membership functions generated with the chromosome with the best fitness in the optimization with 80 % crossover for the input variables are visualized in Fig. 5 and Fig. 6 for the output variable.

Finally, Fig. 7 shows the control space obtained with the membership functions resulting from the genetic algorithm. This image shows the behavior of the voltage variable according to the knowledge base represented by the fuzzy associative memory in Table 1, emphasizing that the voltage applied to the plant is set in the 10 V interval when the error is in the High set, and the flow change is established in the Minimum set.

Conversely, when the governing set of the error is Low, and the flow change is Minimum, the output result is established in the Little set. However, when the electrolyte flow remains stable, the governing output set remains Regular.

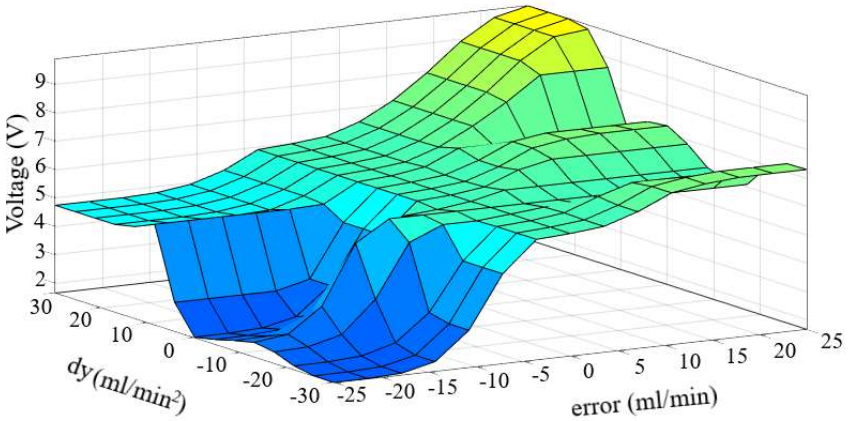


Fig. 7. Control space resulting from the fuzzy regulator obtained with the best-fit chromosome.

4 Conclusions

An evolutionary tuning was developed to establish the optimal parameters of the membership functions of a multiple-input, single-output electrolyte flow controller. Studies were carried out with 60 and 80 % crossover in the population, observing that convergence is obtained earlier when the population that generates offspring is larger. In addition, after performing statistical analysis with the Kolmogorov-Smirnov test, it is demonstrated that a normal distribution behavior and significant differences exist according to the ANOVA test.

On the other hand, regarding the response of the electrolyte flow control system, the tuning of the parameters with the genetic algorithm with 80 % crossover presents a settling time 1.2 s faster than that obtained in the test with 60 % crossover and a difference of 2.6 seconds with the PID controller. Finally, a smaller steady-state error is generated in both evolutionary optimizations as opposed to the classical control obtained with Ziegler Nichols, demonstrating that the heuristics coming from the bio-inspired algorithms are applicable to the optimization of fuzzy control parameters.

References

1. Grzesik W.: Advanced machining processes of metallic materials: Theory, modelling and applications. Elsevier (2017)
2. S. Kalpakjian, S. Schmid: Manufactura, ingeniería y tecnología. Pearson Education (2005)
3. Xu Z. Wang Y.: Electrochemical machining of complex components of aero-engines: Developments, trends and technological advances. Chinese Journal of Aeronautics, vol. 34, no. 2, pp. 28–53 (2021) doi: 10.1016/j.cja.2019.09.016
4. Zaytzev, A., Agafonov, I., Gimaev, N., Moukhoutdinov, R., Belogorsky, A.: Precise pulse electrochemical machining by bipolar current: Aspects of effective technological application.

Journal of Materials Processing Technology, vol. 149, no. 3, pp. 419–425 (2004) doi: 10.1016/j.jmatprotec.2003.10.054
5. Wah, B.: Fuzzy logic control systems. Wiley Encyclopedia of Computer Science and Engineering, pp. 1344-1355 (2008)
6. Trillas, E., Eciolaza, L.: Fuzzy logic: An introductory course for engineering students. Springer International Publishing (2015)
7. Baogang, H., Mann, G., Gosine, R.: New methodology for analytical and optimal design of fuzzy PID controllers. IEEE Transactions on Fuzzy Systems, vol. 7, no. 5, pp. 521–539 (1999) doi: 10.1109/91.797977
8. Charlesworth, B., Charlesworth, D.: Darwin and genetics. Genetics, vol. 183, pp. 757–766 (2009) doi: 10.1534/genetics.109.109991
9. Guo, P., Wang, X., Han, Y.: The enhanced genetic algorithms for the optimization design, de 2010. In: 3rd International Conference on Biomedical Engineering and Informatics, pp. 2990–2994 (2010) doi: 10.1109/BMEI.2010.5639829
10. Ivanova, D., Dejanov, M.: Fuzzy logic control design based on the genetic algorithm for a modular servo system. In: The 17-th International Conference on Electrical Machines, Drives and Power Systems (ELMA), pp. 1–5 (2021) doi: 10.1109/ELMA52514.2021.9503052
11. Murcia, C., Bonilla, G., Melgarejo, M.: Fuzzy classifiers tuning through an adaptive memetic algorithm. IEEE Latin America Transactions, vol. 12, no. 2, pp. 197–204 (2014) doi: 10.1109/TLA.2014.6749538
12. Chang, T., Chang, C.: Genetic algorithm based parameters tuning for the hybrid intelligent controller design for the manipulation of mobile robot. In: IEEE 6th International Conference on Industrial Engineering and Applications, pp. 810–813 (2019) doi: 10.1109/IEA.2019.8715227
13. Shill, P. C., Amin, M. F., Akhan, M., Murase, K.: Optimization of interval type-2 fuzzy logic controller using quantum genetic algorithms. In: IEEE International Conference on Fuzzy Systems, pp. 1–8 (2012) doi: 10.1109/FUZZ-IEEE.2012.6251207.2012
14. Nopalera-Angeles, I.: Algoritmo de control difuso para el ajuste de polarización de un proceso de maquinado electroquímico por pulsos. Tesis de Maestría, Universidad Autónoma del Estado de México (2021)

ISSN 1870-4069

Genetic Programming in Software Engineering

Leslie Loaiza-Meseguer, Angel J. Sánchez-García,
Jorge Octavio Ocharán-Hernández

Universidad Veracruzana,
Facultad de Estadística e Informática,
Mexico

leslielm63@gmail.com, {angesanchez, jocharan}@uv.mx

Abstract. Industry 4.0 has led to automatic optimization of process improvements. Software Engineering is present in all the phases of the Software Development Life Cycle, implying a systematic and disciplined process of development. Nowadays there are optimization problems within the phases and activities of Software Engineering, problems that can be solved with the application of Genetic Programming. The purpose of this Systematic Literature Review is to analyze the current state of the application of genetic programming in Software Engineering by collecting the phases and activities of software development where genetic programming has been used and summarizing the advantages of using this technique. Thanks to this research work we found the way in which genetic programming has been applied previously and the advantages that its application has both in functional and non-functional properties. In addition, the utility that it has for a software engineer to use this technique as an automation tool in the process of software development was found.

Keywords: Genetic programming, software engineering, systematic literature review, optimization.

1 Introduction

Industry 4.0 seeks to improve processes and products through the incorporation of new technologies, cloud computing, the Internet of Things and Artificial Intelligence, among others. Software Engineering, for other hand, seeks to develop computer systems through a systematic, disciplined and orderly process in order to obtain a quality product and reduce the number of defects.

This implies that Software Engineering is present in all the phases of the life cycle of a software project [1]. Nowadays there are optimization problems within the phases and activities of Software Engineering that need to be solved since, during the construction of a project, several factors can be found that negatively influence its performance, production time, and reliability, among other aspects [2].

Genetic programming is applicable and effective for a wide variety of problems that arise in a wide variety of fields, mainly for the development of computer programs that perform a user-defined task. In addition, this technique is able to take advantage of the exponential increase in available computational power and solve optimization problems

Table 1. Research questions.

Quesition	Motivation
RQ1.- In which phases of software development has genetic programming been used?	The purpose of this question is to know the phases of software development in which genetic programming has been used to identify promising areas of the use of this technique or its variants.
RQ2.- In which activities of the software development phases has genetic programming been used?	It is important to identify which are the specific activities of each of the Software Engineering phases where genetic programming has been applied to promote improvements in software engineers.
RQ3.- What are the advantages of using genetic programming?	It is important to know about the benefits of applying genetic programming and why to use it in the different Software Engineering activities.

Table 2. Keywords and synonyms identified.

Concept	Synonyms
Software Engineering	
Genetic programming	GP

[3]. In Software Engineering, genetic programming has been used to represent code structures. It is reported that there are problems in the construction phase that have been addressed with optimization algorithms [4], specifically with genetic programming for code refactoring [5].

However, genetic programming is not limited to the coding phase, as it has impacted activities such as Software reliability [5], code repair [6], defect prediction [7], among others. With the above, it can be seen that the application of genetic programming supports Software Engineers, providing them with tools that help them to increase the efficiency of their work.

This paper is organized as follows: Section 2 describes the background as well as related work. In Section 3, the method used to carry out this research work is detailed. Section 4 presents the results obtained. Finally, section 5 draws the conclusions and future work.

2 Background and Related Work

Software engineering activities employ economic and human resources and involve the investment of time. Whether a software development project is successful or not depends entirely on the human factor[2], which is found in different phases of software development, such as design, construction and maintenance.

As a consequence of these factors, artificial intelligence techniques have been used to help reduce time, costs and human errors. Recent works have shown that Artificial Intelligence can bring benefits in each of the phases of software development, for example, requirements analysis [8], design [9], coding and testing[10].

It is reported that there are problems in the construction phase that have been addressed with optimization algorithms [4] specifically with genetic programming for

Table 3. Data source.

Database	Website
IEEE Xplore	https://ieeexplore.ieee.org/Xplore/home.jsp
Science direct	https://www.sciencedirect.com/
ACM	https://dl.acm.org/
Springer Link	https://link.springer.com/

Table 4. Inclusion criteria.

ID	Description
IC1	Studies with full access.
IC2	Studies published between 2017 and 2022
IC3	Studies that in the title or abstract allude to any of the phases or activities of software engineering.
IC4	Studies that contain in the abstract indications of answering at least one research question.

code refactoring [5]. Genetic programming is an extension of the traditional genetic algorithm in which each individual in the population is represented as a program variant (patched program).

The program variant is generated using one of the operations of the genetic algorithm: mutation and crossover. The acceptability of each variant is calculated through a user-defined fitness function.

These program variants that obtain high fitness scores are selected for the next evolution. The evolution process will continue again and again until a valid patch is found [6].

In a manual search of related work, no Systematic Literature Review on the application of genetic programming in Software Engineering was identified. For this reason, the main objective of this research is to describe the current state of the use of genetic programming in each of the phases of the software development life cycle, emphasizing the benefits for (although not limited to) software engineers, software developers and testers.

3 Research Method

The method used to carry out this systematic review of the literature is based on the guidelines proposed by Kitchenham & Charters [11], which are described below.

3.1 Research Questions

The research questions that guided this systematic review are shown in Table 1.

Table 5. Exclusion Criteria.

ID	Description
EC1	Studies in a language other than English.
EC2	Studies that are book chapters, presentations, abstracts or technical reports.
EC3	Repeated or duplicated studies.

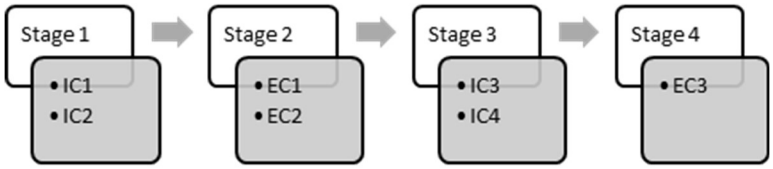


Fig. 1. Primary study selection procedure.

3.2 Search Strategy and Data Sources

The terms used to search for the primary studies are defined in Table 2. Being an exploratory study, each phase of software development (such as requirements, design, coding or testing) was not placed in the search string.

In addition, the term "Software Engineering" was added, which implies a development process, instead of putting only the term "software" since studies referring to the use of software to test genetic programming in different areas could be included. The search string used is based on the search terms defined above and is made up as follows:

"Software Engineering" AND "Genetic programming"

Table 3 shows the databases that were selected for the search of the primary studies, as well as their e-mail addresses.

3.3 Selection of Primary Studies

The inclusion and exclusion criteria described in Table 4 and 5 are intended to determine which studies will be included or excluded for the elaboration of this study.

3.4 Selection Procedure

The primary study selection procedure consists of four stages. Fig. 1 shows in detail the primary study selection criteria previously defined in section 3.3 that are applied in each of these stages. Table 6 shows in detail the results of each database during the 4 stages.

The list of references of the 41 primary studies selected can be found in [12]. The template used to extract data from each primary study can be found in [13]. The questions defined in order to evaluate the quality of primary studies can be found in [14].

Table 6. Application of inclusion and exclusion criteria by stage.

Stage	IEEE Xplore	ACM	SpringerLink	ScienceDirect	Total
Initial search	182	613	2,741	476	4,012
Stage 1	28	235	226	31	520
Stage 2	25	196	190	17	428
Stage 3	12	18	14	2	46
Stage 4	12	13	14	2	41

4 Results

As a result of the application of stage 4 of the primary study search selection process, of the 41 resulting studies were identified. The most came from SpringerLink (34%), followed by ACM (32%), IEEE Xplore (29%) and ScienceDirect (5%). Of these selected studies, 59% correspond to articles published in journals, while 41% are conference papers, as it is shown in Fig. 2.

The distribution of primary studies by year of publication was also identified, with the majority of studies coming from 2017 and 2021, as it is shown in Fig. 3.

4.1 RQ1.- In Which Phases of Software Development has Genetic Programming Been Used?

As can be seen in Fig. 4, the phase that has had the most applications of genetic programming, is the construction phase, occupying 68% of the total selected studies.

It was found that the tree structure used by genetic programming to represent its individuals (computer programs) is very useful for the construction of software, since it allows the creation of a new code from an existing code [15]; thus, removing branches from one tree to insert them into another [16], which promotes the improvement of both functional and non-functional properties [17], automatic code generation[18], as well as code reuse and restructuring[19].

Genetic programming was found to be a tool that facilitates pattern identification [20], this property allows us to identify code smells [21], locate faults [22] and patch generation [23]; the latter enables automatic program repair [24]. On the other hand, the Testing phase is mentioned in 20% of the articles.

Wei et al. [25] again points out the ability of genetic programming for pattern identification, which, according to their study, allows the identification of the worst-case scenario in the execution of a software, as well as the detection of vulnerabilities and performance errors.

Other authors propose that the application of genetic programming allows the automation of black box testing [26] and the evaluation of graphical interfaces [27]. Regarding the Design phase, only 7% of the articles were found to mention it. It was found that genetic programming helps the automation of both prototype generation [28] and modeling of software product lines [29].

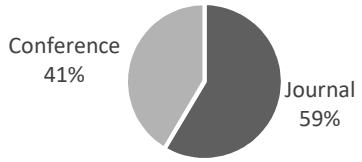


Fig. 2. Selected primary studies by publication type.

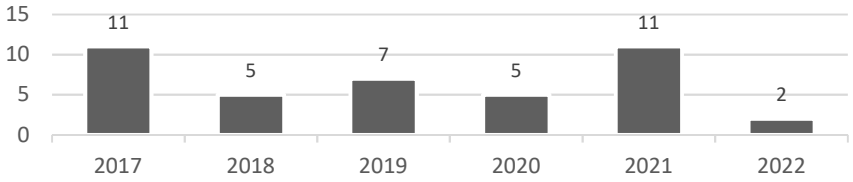


Fig. 3. Selected primary studies by year.

Finally, the Planning phase was found to be present in 3% of the articles and the Maintenance phase in 2%, indicating a lack of information on the areas of opportunity for genetic programming in these phases.

4.2 RQ2.- In Which Activities of the Software Development Phases Has Genetic Programming Been Used?

Twenty-seven papers that mention activities of software development phases involving the application of genetic programming were found. In the Planning phase, by applying genetic programming in the restructuring of plans, initializing the population of individuals with existing plans, we can reuse their information to carry out the generation of new plans [30].

In the Design phase it was found that 67% of the items correspond to prototype generation. This activity is achieved by automating the definition of basic elements and the way to combine them, to be subsequently composed and tested with real users and thus find the optimized compositions [31].

The 33% of the articles refer to the modeling of Software Product Lines (SPL). The automatic generation of the generic models used by this activity is possible by means of an initial population of these and the calculation of the set of valid characteristics for each one, to subsequently apply genetic programming to them [29].

In the Construction phase we were able to identify that 32% of the items correspond to the activity of automatic program repair. This activity aims to generate error repairs without human intervention, without the need for special instrumentation or annotations in the source code [32].

This application searches and generates modifications from an abstract syntax tree that can patch a bug in the underlying program and creates new program variants by mutation and crossover [15].

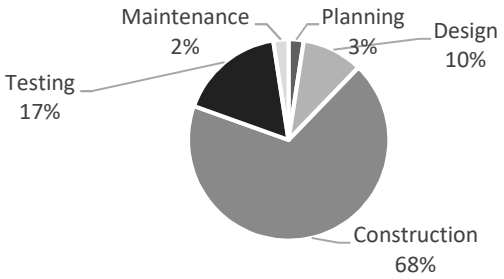


Fig. 4. Selected primary studies by phase.

The 32% of the articles mention automatic coding of programs, for this activity it is important to mention that one of the most relevant applications of genetic programming is the technique called program synthesis, which is able to automatize the coding of programs by automatically generating source code in a programming language, that maintains the constraints of a predefined specification [32].

It was determined that 31% of the articles correspond to the identification of bugs in the software, which is achieved through the identification of recurrences [24] and sequence-by-sequence learning [15].

These properties are useful in code smells identification [21] and fault localization [17]. Finally, 5% of the articles mention that genetic programming supports code restructuring by identifying patterns within the code, in order to subsequently optimize its fragments and reuse them to create new code [19].

In the testing phase, it was found that 67% of the articles talk about black box testing, to which the application of genetic programming is useful to discover local variables, actions performed on output variables, counting loops and while loops, due to the ability of genetic programming to discover a functional relationship between data features and to group them into categories [26].

Also, the application of genetic programming in black box testing is useful when identifying worst-case execution as well as vulnerabilities in programs by identifying patterns in data inputs [25].

The 33% of the articles allude to interface evaluation, where genetic programming allows the automatic generation of rules to evaluate their quality, providing previously defined quality metrics, context criteria and list of possible types of problems, taking advantage of the principle of genetic programming where individuals adapt to their environment through mutation and crossover [27].

4.3 RQ3.- What are the Advantages of Using Genetic Programming?

Based on the answers to questions RQ1 and RQ2, it was found that thanks to the mutation and crossover operators, the principles of biological evolution on which genetic programming is based and its ability to identify patterns, there are numerous advantages in the different phases and activities of software development.

The use of genetic programming in the Planning phase, serves for the restructuring of plans, reducing the costs of operating in complex environments of change and uncertainty, by adapting autonomously to change in the pursuit of its quality objectives

[30]. In the Design phase, prototyping [31] and modeling of software product lines [29] can be automated using genetic programming, greatly improving the performance of these activities [28].

In the Construction phase, the application of this technique helps in the automation of different activities such as program repair [32], bug identification [24] and code restructuring [19]. Obtaining the improvement of both functional and non-functional properties, such as code size, execution time or memory consumption [17].

Furthermore, genetic programming is a technique that has great flexibility since it offers the possibility of handling a large number of individuals and of reworking the solutions obtained by relaunching a new evolution from one or more previously obtained solutions, so that, with its application, the activity of evaluating graphical interfaces can be automated and thus optimize the process involved [17].

All this together helps a software engineer to do his job efficiently since it eliminates the manual part of his work and increases the quality of his results.

5 Conclusions and Future Work

This paper identified the phases and activities of software development in which genetic programming has been applied, as well as the advantages of using it. A systematic Literature Review was carried out where the selection process of primary studies was divided into 4 parts where, after applying the previously defined selection criteria, 41 primary studies were obtained as a result.

Subsequently, after carrying out a preliminary data synthesis, the primary studies were classified into 5 development phases: Planning, Design, Construction, Testing and Maintenance. The software development phase where genetic programming proved to have more applications is the Construction phase with 28 articles, followed by the Testing phase with 8 articles, the Design phase with 3 articles and the Planning and Maintenance phases with 1 article each.

With respect to the research questions, thanks to the data synthesis, it was found the way in which genetic programming has been previously applied, the advantages of its application in both functional and non-functional properties, in addition to the usefulness for a software engineer to use this technique as an automation tool in the different processes that exist at the time of software development. As a result, the objectives of the research work were achieved.

It was found that genetic programming has a great relationship with well-established areas, for example, Program synthesis, which has a strong impact on new fields such as Genetic improvement. This field of science uses genetic programming to correct bugs in software and improve both functional and non-functional software requirements [19]. Therefore, as future work, we will seek to identify the applications and advantages of these areas.

References

1. Boehm, B.: Software engineering. IEEE Transactions on Computers, pp. 1226–1241 (1976) doi: 10.1109/TC.1976.1674590

2. Yanyan, Z, Renzuo, X.: The basic research of human factor analysis based on knowledge in software engineering. In: 2008 International Conference on Computer Science and Software Engineering, pp. 1302–1305 (2008) doi: 10.1109/CSSE.2008.219
3. Koza, J. R.: Genetic programming: On the programming of computers by means of natural selection, MIT press (1992)
4. Robles-Aguilar, A, Ocharán-Hernández, J. O., Sánchez-Garcia, A. J., Limon, X.: Software design and artificial intelligence: A systematic mapping study. In: 2021 9th International Conference in Software Engineering Research and Innovation (CONISOFT), pp. 132–141 (2021) doi: 10.1109/CONISOFT52520.2021.00028
5. Chen, H, Zhang, Y, Zhao, J.: Improved genetic programming model for software reliability. In: 2009 International Asia Symposium on Intelligent Interaction and Affective Computing, pp. 164–167 (2009) doi: 10.1109/ASIA.2009.38
6. Qi, Y., Mao, X., Lei, Y., Dai, Z., Wang, C.: Does genetic programming work well on automated program repair? In: 2013 International Conference on Computational and Information Science, pp. 1875–1878 (2013) doi: 10.1109/ICCIS.2013.490
7. Rathore, S. S., Kuamr, S.: Comparative analysis of neural network and genetic programming for number of software faults prediction. In: 2015 National Conference on Recent Advances in Electronics and Computer Engineering (RAECE), pp. 328–332 (2015) doi: 10.1109/ RAECE.2015.7510216
8. Ernst, N. A., Gorton, I.: Using AI to model quality attribute tradeoffs. In: 2014 IEEE 1st International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), pp. 51–52 (2014) doi: 10.1109/AIRE.2014.6894856
9. Wangoo, D. P.: Artificial intelligence techniques in software engineering for automated software reuse and design. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1–4 (2018) doi: 10.1109/CCAA.20 18.8777584
10. Xie, T.: The synergy of human and artificial intelligence in software engineering. In: 2013 2nd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), pp. 4–6 (2013) doi: 10.1109/RAISE.2013.6615197
11. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University Joint Report (2007)
12. Appendix A Primary studies references: https://docs.google.com/document/d/11uCIfSXL-hxb1xVBhpcbCOF1T0rUyC99TBMfdck9R5c/edit?usp=sharing
13. Appendix B Data extraction template: https://docs.google.com/document/d/1JQ4 x7up_ZlkXBol26z500ff-wK4DlH14IWh274JBJp8/edit?usp=sharing
14. Appendix C Quality assessment questions: https://docs.google.com/document/d/1RQuy Ztujhr0wHsWES8dR2nswFO25pIjLozUTfKcxZao/edit?usp=sharing
15. Li, D., Wong, W. E., Jian, M., Geng, Y., Chau, M.: Improving search-based automatic program repair with neural machine translation. IEEE Access, vol. 10, pp. 51167–51175 (2022) doi: 10.1109/ACCESS.2022.3164780
16. Langdon, W. B., Lam, B. Y., Modat M., Petke, J., Harman, M.: Genetic improvement of GPU software. Genetic Programming Evolvable Machines, vol. 18, pp. 5–44. (2017) doi: 10.1007/s10710-016-9273-9
17. Sohn, J., Yoo, S.: Empirical evaluation of fault localization using code and change metrics. IEEE Transactions on Software Engineering, vol. 47, no. 8, pp. 1605–1625 (2021) doi: 10.1109/TSE.2019.2930977
18. Miller, J. F.: Cartesian genetic programming: Its status and future. Genetic Programming Evolvable Machines, vol. 21, pp. 129–168 (2020) doi: 10.1007/s10710-019-09360-6
19. Krauss, O.: Genetic improvement in code interpreters and compilers. In: Proceedings Companion of the 2017 ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity, pp. 7–9 (2017) doi: 10.1145/3135932.3135934

20. Huppe, S., Saied, M. A., Sahraoui, H.: Mining complex temporal API usage patterns: An evolutionary approach. In: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), pp. 274–276 (2017) doi: 10.1109/ICSE-C.2017.147
21. Kessentini, M., Ouni, A.: Detecting android smells using multi-objective genetic programming. In: 2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILESoft), pp. 122–132 (2017) doi: 10.1109/MOBILE Soft.2017.29
22. Kim, Y., Mun, S., Yoo, S., Kim, M.: Precise learn-to-rank fault localization using dynamic and static features of target programs. ACM Transactions on Software Engineering and Methodology, vol. 28, no. 4, pp. 1–34 (2019) doi: 10.1145/3345628
23. Cao, H., Liu, F., Shi, J., Chu, Y., Deng, M.: Automated repair of Java programs with random search via code similarity. In: 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 470–477 (2021) doi: 10.1109/Q RS-C55045.2021.00075
24. Yuan, Y., Banzhaf, W.: Toward better evolutionary program repair. ACM Transactions on Software Engineering and Methodology, vol. 29, pp. 1–53 (2020) doi: 10.1145/3360004
25. Wei, J., Chen, J., Feng, Y., Ferles, K., Dillig, I.: Singularity: Pattern fuzzing for worst case complexity. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 213–223 (2018) doi: 10.1145/3236024.3236039
26. Drusinsky, D.: Reverse engineering concurrent UML state machines using black box testing and genetic programming. Innovations in Systems Software Engineering, vol. 13, pp. 117–128 (2017) doi: 10.1007/s11334-017-0299-9
27. Ines, G., Makram, S., Mabrouka, C., Mourad, A.: Evaluation of mobile interfaces as an optimization problem. Procedia Computer Science, vol. 112, pp. 235–248 (2017) doi: 10.1016/j.procs.2017.08.234
28. Valencia-Ramírez, J. M., Graff, M., Escalante, H. J., Cerda-Jacobo, J.: An iterative genetic programming approach to prototype generation. Genetic Programming Evolvable Machines, vol. 18, pp. 123–147 (2017) doi: 10.1007/s10710-016-9279-3
29. Vescan, A., Pintea, A., Linsbauer, L., Egyed, A.: Genetic programming for feature model synthesis: A replication study. Empirical Software Engineering, vol. 26, no. 58 (2021) doi: 10.1007/s10664-021-09947-7
30. Kinneer, C., Coker, Z., Wang, J., Garlan, D., Le-Goues, C.: Managing uncertainty in self-adaptive systems with plan reuse and stochastic search. In: Proceedings of the 13th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS), pp. 40–50 (2018) doi: 10.1145/3194133.3194145
31. Salem, P: User interface optimization using genetic programming with an application to landing pages. In: Proceedings of the ACM Human-Computer Interaction, vol. 1, pp. 1–17 (2017) doi: 10.1145/3099583
32. Sobania, D., Rothlauf F.: A generalizability measure for program synthesis with genetic programming. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 822–829 (2021) doi: 10.1145/3449639.3459305

ISSN 1870-4069

The P-Median as a Problem of Clustering

María Beatriz Bernábe-Loranca[1], Carmen Cerón-Garnica[1],
Hugo Rodríguez-Cortes[2], Rogelio González-Velázquez[1]

[1] Benemérita Universidad Autónoma de Puebla,
México.

[2] Instituto Politécnico Nacional,
Centro de Investigación y de Estudios Avanzados,
México

{maria.bernabe, carmen.ceron,
rogelio.gonzalez}@correo.buap.mx,
hrodriguez@cinvestav.mx

Abstract. If a problem of Area Design can be modeled under conditions of classic Partitioning, then the problem is often called Territorial Clustering (TC) since the computational solution uses algorithmic techniques of portioning clustering. Under these characteristics, Partitioning can be seen as a methodology of support to solve problems of P-median and territory design type. Partitioning in the solution of territorial problems consists in to cluster small geographic areas called basic unity in a given number of bigger groups named territories. Such description requires of a mathematical model for its expression, and one of the useful definitions in the model is the definition of discrete partition. Is in this point where the following work is situated: it is proposed to show that the definition of Zones Design (ZD) and partitioning share some properties in their definition, both in the restrictions and in objective function. Likewise, the transformation of the P-median problem model of integer binary problem to a combinatorial optimization model is posed.

Keywords: Territory design, integer-binary model, combinatorial model, P- median.

1 Introduction

The problems of territorial clustering have applications in the determination of political and scholar districts, installation of social and emergency services, commercial territories, etc. In various works, geographic criteria are used as adequacy measures of solutions, for example, that a territory is geographically compact, and it is often used as compactness measure the sum of distances between the basic unities and the centroid to which they are assigned, thus modeling the problem as the P-median.

The modeling of clustering problems adjusted as the P-median motivate the study of instances of big scale, as an example, the problems of P-median defined in graphs G = (V,A) with |A| >= 360,000 are difficult to solve with commercial software intended for

problems of Mixed Integer Programming, then additional strategies are needed such as the metaheuristics to solve the problem.

However, the objective in this document is centered in to present the conditions of equivalence between the P-median and the clustering problem. On the other hand, the transformation of the model of the P-median is proposed from an integer-binary to a combinatorial.

2 Preliminaries

Definition 1. Let the initial set of unities of area be $X = \{x_1, x_2, \dots, x_n\}$ where the *ith* unity of area is x_i. The number of zones is denoted by k and Z_i is the set of all unities of area that belong to the zone Z_i [1]. Then:

$$Z_i \neq \emptyset \quad \text{for i} = 1 \dots k, \tag{1}$$

$$Z_i \cap Z_j = \emptyset \ \text{ for i} \neq \text{j y}, \tag{2}$$

$$\cup_{i=1}^{k} Z_i = X. \tag{3}$$

Therefore, (1), (2) and (3) constitute the set of "equivalent restrictions" both in the clustering and in zones design.

It is stablished that these criteria comply indirectly with at least the geometric compactness, although in several computational solved problems it has been graphically observed that compactness and connectedness are satisfied.

In other cases, it has been identified in maps that the contiguity is also fulfilled, but this visual observation is valid only for some applications such as the P-median and it is not guaranteed that the contiguity is reached, even some authors have already warned that the inclusion of contiguity represents a more complex problem [2].

Zones design, territorial design, design of territory in a pragmatic sense, share the same semantic, also the terms of zoning or regionalization are [3-5].

Now, considering the definition 1 and observing the following definition 2:

Definition 2. Let a partition of a set A of n elements in k pairs, be a family of k non-empty disjoint subsets of A, such that their union is the same set A, thus $A_1, \dots, A_k$, are the subsets of the partition, then it is fulfilled:

2.1 Problem of Territorial Clustering as a Problem of the P-Median

An application of Territorial Design (TD) that can be employed as methodology when criteria of compactness and connectedness are treated, are the location-allocation models. When talking about model, it refers to the mathematical representation of the problem and it is translated to a methodology always that points of the model are respected and adapted to a series of algorithmic instructions.

This leads to the election of adequate techniques that answer to the model, such that candidate algorithms are identified for its solution and consequently, evaluate them to give answer to the model.

However, any variant in the model that does not change the meaning, implies a strategy reasonably distinct, but that leads to the same solution, it is said that the model

of the basic P-median is the same in all the literature, although the methodology and proposal of algorithm that are suggested for its solution is different but searching the same result.

For example, when different metaheuristics are used to answer the P-median in the combinatorial model, the solutions are not exactly the same since they are approximated, but always trying that such approximations get close to the optimum, or in the best of the cases to be the exact value. To value the implementations, the instances of the P-median test are found on the website OR-Library [6].

The P-median can be stablished in terms of graphs as follows:

Let G = (*V*, E) be a graph not aimed where V is the set of n vertices and E is the set of edges with a weight associated that can be the distance between the vertices $d_{ij} = d(v_i, v_j)$ for all $i, j = 1, \dots, n$ according to certain metrics, then with the distances a symmetric matrix is formed, then $V_p \subseteq V$ must be found such that $|V_p| = p$, where p can be either variable or fixed, and that the sum of the shortest distances of the vertices in $\{V - V_p\}$ to its nearest vertex in V_p is reduced to the minimum.

Particularly, the P-median is useful in these conditions independently of the operations research approach (integer-binary) or the combinatorial problem (tackled with metaheuristics).

The problem of the P-median has been widely studied in literature. In [7] can be found an excellent revision of methodologies to solve the P-median of approximated form and exact or through graphs.

Geographically, P-median is observed to partition the territory, then it can be said that just for this characteristic, the P-media belongs to problems of Territorial Design, which have diverse applications such as the identification of political districts, social services installation, commercial territories, location-allocation, etc. [3-5].

In lots of works, geographic criteria of adequacy measures of the solutions are utilized. The criteria commonly used are compactness, connectedness and contiguity. According to Kalcsis, a territory is geographically compact if it has a round approximated form and it is not distorted, but there is no rigorous definition of the concept [3].

3 Analysis of the Clustering Problem as One of Optimization

Let the clustering be a problem of partitioning grouping, then given a set of n objects denoted by $X = \{x_1, x_2, \dots, x_n\}$ in that $x_i \in R^D$ let k be a positive integer known a priori, the clustering problem consists in finding a partition:

$P = \{C_1, C_2, \dots, C_k\}$ of X, being C_j a conglomerate (group) conformed by similar objects, satisfying an objective function f: $R^D \to R$, and the conditions:

$$C_i \cap C_j = \emptyset \text{ for } i \neq j \text{ and } \bigcup_{j=1}^{k} C_j = X$$

To measure the similarity between two objects x_a and x_b, it is used a function of distance denoted by $d(x_a, x_b)$, being the Euclidean distance the most popular to measure the similarity. Thus, the distance between two different elements:

$$x_i = (x_{i1}, \dots, x_{iD}) \quad and \quad x_j = (x_{j1}, \dots, x_{jD})$$

is $d(x_i, x_j) = \sqrt{\sum_{l-1}^{D}(x_{il} - x_{jl})^2}$.

The objects of a conglomerate are similar when the distances between them is minimal; this allows to formulate the objective function f, as:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \overline{x_j}). \quad (1)$$

That is to say, it is desired to minimize (1); where $\overline{x_j}$, known as representative element of the conglomerate (group), is the mean of the elements of the conglomerate:

$$C_j, \overline{x_j} = \frac{1}{|c_j|} \sum x_i \in C_j. \quad (2)$$

And it corresponds to the center of the conglomerate. Under these characteristics, clustering is a problem of combinatorial optimization, and has been demonstrated that is NP-difficult.

3.1 P-Median Classic Approach

The P-median consists in, given a set of points (or location of consumers) and a matrix of distances (or costs) between all and each of points, choosing p points (or location of installations) with the purpose of minimizing the sum of each of the distances of all points to the "nearest chosen point".

In 1970 ReVelle and Swain presented the first formulation of integer programming for the p-median problem cited by Church in 2003 [8-12]. In general, the p-median problem can be expressed mathematically as a problem of discrete optimization.

First, the matrix of distances is denoted as d_{ij}, that expresses the distance between the potential points of location i and the points of demand j. The binary variable x_{ij} corresponds to the allocation or not allocation of the demand points j to the installation i. The binary variable y_i indicates that an installation is stablished in the point i or not. The approach as an integer binary problem is the following form:

Let $x_{ij} = \begin{cases} 1 & \text{If the point } j \text{ is assigned to the point } i \\ 0 & \text{in other case;} \end{cases}$.

and $y_i = \begin{cases} 1 & \text{If in the point } i \text{ is located in an installation} \\ 0 & \text{in other case;} \end{cases}$

$$min \quad Z = \sum_{i=1}^{k} \sum_{j=1}^{n} d_{ij} x_{ij}\,, \quad (a)$$

$$\text{subject to} \quad \sum_{i=1}^{k} x_{ij} = 1 \quad \forall j = 1,2,\ldots,n; \quad (b)$$

$$\sum_{j=1}^{n} x_{ij} \leq n y_i \quad \forall i = 1,2,\ldots,k; \quad (c)$$

$$\sum_{i=1}^{k} y_i = p\,, \quad (d)$$

where k is the number of potential vertices where a median, generally$k = n$, can be localized, p is the fixed number of required medians. The equation (a) is the objective function that minimizes the distance of the system, the restriction (b) stablishes that each demand point can only be allocated in an installation, the restriction (c) stablishes that the assignation of demand points to each one of the installations or medians and finally the restriction (d) guarantees that between k potential points of location that choose exactly p.

Test. We define, $u_{jr} = \begin{cases} 1 & if\ \ r\ is\ assigned\ to\ \ j \\ 0 & in\ other\ case \end{cases}$

Since as each point belongs to a solely group, it is satisfied $\sum_{j=1}^{k} u_{jr} = 1$

Moreover, $\sum_{x_i \in C_j} d(x_i, \overline{x_j}) = \sum_{r=1}^{n} u_{jr} d(x_i, \overline{x_j})$

Then $\sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, \overline{x_j}) = \sum_{j=1}^{k} \sum_{r=1}^{n} u_{jr} d(x_i, \overline{x_j})$

If the problem is interpreted as potential points of location j and the demand points r. The binary variable u_{jr} corresponds to the allocation of demand points r to the installation j or not. The variable $\overline{x_j}$ can be interpreted as a installation that is stablished in the group C_j:

$$y_j = \begin{cases} 1 & in\ the\ installation\ \ \overline{x_j} \\ 0 & in\ other\ case \end{cases}$$

Then it is satisfied:

$$\sum_{j=1}^{k} y_j = k \quad and \quad \sum_{r=1}^{n} u_{jr} = n y_j$$

In conclusion, all hypothesis of the model of the P-median and vice versa are fulfilled.

4 Transformation of the Model of the P-Median Problem of Integer Binary Programming to Combinatorial Optimization

Is of interest to show the transformation of the model of integer-binary programming of the P-median to one of combinatorial optimization.

To implement approximation algorithms, in the search of solutions to the NP-hard problems such as the P-Median is, it must be approached as a combinatorial optimization problem [13].

The problem of the P-median is correct in the solution of problems of localization, of territorial design, set partitioning, of cluster design, among others and it is necessary to construct methods of search of solutions such as the metaheuristics [11].

The applications of the localization problems are responsible of solving the location of one or several installations, in this way optimizing one or various objectives, such is the case of transport cost, client service, market partition, etc. The study of localization problems involves a lot of fields of knowledge as operations research, administration science, industrial engineering, computer science, urban planning and different related areas. Among the applications to the design of logistic networks, there is territorial

design, location of warehouses, production plants, assembly plants, hospitals, fire stations, police stations, schools, etc.

To show the transformation of the mathematical model of the problem de la P-median as one of programming integer binary to one of combinatorial optimization, the following nomenclature is given:

PPM= problem of the p-median,

PIBP=problem of integer binary programming,

PCO= problem of combinatorial optimization.

First, it is important an introduction to the model of combinatorial optimization of P-median. Let be the problem of the P-Median (PPM) in the following way:

Given a set of n vertices of a graph in the plane denoted by$V = \{1, 2, \dots, n\}$, $|V| = n$, the objective is to find a subset of vertices $L \subseteq V$, with $|\mathrm{L}| = p$ of potential locations of the medians such that the total mean distance of the design is the minimum. It is said that this approach of is of type Problem of Combinatorial Optimization (PCO) since the space of feasibility Ω of PPM are all the subsets L of cardinality p of a set of cardinality n, with $p < n$, this is:

$$|\Omega| = \frac{n!}{p!\,(n-p)!} = \binom{n}{p}$$

In general, a instance for the PPM is denoted by PPM(n, D, p), where n is the cardinality of the set V, p is the cardinality of the obtained subsets y $D = (d_{ij})$ is the matrix of distances among all pair of vertices of V.

Second, let be the problem of the P-median seen as an integer-binary model, then the mathematical model of the PIBP of the PPM is developed as follows:

The PPM about a graph can be expressed mathematically as an integer-binary programming problem (PIBP) of the form [8]:

$$\text{Let } x_{ij} = \begin{cases} 1 & \text{if the vertex } j \text{ is assigned to the vertex } i \\ 0 & \text{in any other case} \end{cases}$$

$$\text{Sea } y_i = \begin{cases} 1 & \text{if the vertex } i \text{ is a median} \\ 0 & \text{in any other case} \end{cases}$$

$$Min\ Z = \sum_{i=1}^{k} \sum_{j=1}^{n} d_{ij} x_{ij}, \tag{1}$$

$$\text{subject to } \sum_{i=1}^{k} x_{ij} = 1 \quad \forall\, j = 1, \dots, n \quad , \tag{2}$$

$$\sum_{j=1}^{n} x_{ij} \leq n\, y_i \quad \forall\, i = 1, \dots, k \quad , \tag{3}$$

$$\sum_{i=1}^{k} y_i = p \ . \tag{4}$$

Table 1. Binary solutions.

Binary Solution	Median	Group
$x_{61} = 1, x_{62} = 1, x_{65} = 1, x_{69} = 1, x_{611} = 1$	6	6, 1, 2, 5, 9, 11
$x_{73} = 1, x_{74} = 1, x_{78} = 1, x_{710} = 1$	7	7, 3, 4, 8, 10
$x_{1512} = 1, x_{1514} = 1, x_{1517} = 1$	15	15, 12, 14, 17
$x_{1316} = 1$	13	13, 16

Table 2. Results for the OR-Library P-Median instances (Part 1).

Instance	VNS		SA		VNS-BIO	
	Cost		Cost	T	Cost	T
1	5819	103	6209	28	5884	0.55
2	4341	180	4646	70	4601	7.872
3	4467	180	4785	47	4870	7.793
4	3380	250	3693	93	3744	2.877
5	1664	360	1820	119	1771	2.749
6	7917	330	8349	49	8236	2.427
7	5952	540	6446	104	6415	3.1
8	5204	1003	5536	174	5507	2.647
9	3385	1860	3626	286	3432	2.134
10	1700	1980	1850	907	1844	2.662
11	7803	720	8346	149	7968	9.762
12	7200	900	7717	298	7485	9.615
13	5126	1860	5475	692	5444	9.877
14	3823	1440	3992	71	3907	10.403
15	2464	240	2558	1721	2482	10.354
16	8423	300	8958	220	8736	21.732
17	7651	540	8197	322	7905	20.896
18	5821	2700	6038	505	5935	22.986
19	3747	2760	3881	2036	3780	24.238
20	2647	2040	2755	1909	2707	24.865
21	9557	240	10231	102	9794	43.587
22	9433	300	9802	170	9744	38.277
23	5645	3600	5941	839	5827	44.257
24	3974	600	4065	1440	3988	46.955
25	2726	720	2852	240	2810	51.757

26	10312	60	10869	14	10458	64.285
27	9065	120	9511	25	9159	80.152
28	5664	480	5799	141	5768	75.77
29	4114	780	4176	70	4132	86.531
30	2960	1320	3058	47	2996	122.48
31	10528	70	11157	93	10739	105.61
32	10383	120	10818	119	10606	34.87
33	6007	720	6166	49	6102	43.789
34	4193	1500	4286	104	4191	53.045
35	11037	120	11698	174	11180	47.757
36	9994	180	11544	250	11154	60.261
37	6460	1620	6715	50	6602	62.387
38	11725	180	12252	10	11678	68.799
39	10570	300	11017	25	10599	62.698
40	6632	2460	6803	40	6664	72.65
32	10383	120	10818	119	10606	34.87
33	6007	720	6166	49	6102	43.789
34	4193	1500	4286	104	4191	53.045
35	11037	120	11698	174	11180	47.757
36	9994	180	11544	250	11154	60.261
37	6460	1620	6715	50	6602	62.387
38	11725	180	12252	10	11678	68.799
39	10570	300	11017	25	10599	62.698
40	6632	2460	6803	40	6664	72.65

The equation (1) is the objective function with variables of decision x_{ij} and y_i binary. The number of potential vertices where a median can be located is k and generally $k = n$. The fixed number of required medians is p. The restrictions (2) guarantee that every vertex has associated a median. The restrictions (3) indicate the distribution of the vertices to the medians. The (4) determines the number of medians.

4.1 The Proceeding for the Transformation through the Analysis of a Case

The PPM as problem of integer binary programming (PIBP) has a space of feasibility of exponential type 2^n and as a POC is $\binom{n}{p}$

Then a transformation is $T: 2^n \rightarrow \binom{n}{p}$

Considering a case of a graph of $n = 17$ vertices, that is to say, $V = \{1, 2, \ldots, 17\}$ illustrated in Fig. 1, whose solution was determined by the PIBP model that is observed in Fig. 2. The results with a cost C determined by the objective function of the equation (1) are represented in the following figures.

The solution is shown in Table 1 and the binary solution determines the medians in $y_6 = 1, y_7 = 1, y_{15} = 1, y_{13} = 1$.

The subsequent combinatorial solution obtained with an enumerative algorithm of Partitioning, are the "centroids" $L = \{6, 7, 13, 15\}$. Said solution has a cost C obtained of the sum of the sums of the distances of the "medians", now expressed as centroids to their associated vertices, that is to say:

C = d(6,1)+ d(6,2)+d(6,5)+ d(6,9)+ d(6,11)+ d(7,3)+ d(7,4)+ d(7,8)+ d(7,10)+ d(15,12)+ d(15,14)+ d(15,17)+ d(13,16).

The test instance for the combinatorial problem of the P-Median $(17, D_{17x1}, 4)$ has shown that the equivalency between the solutions of PIBP and PCO for the PPM.

5 Application

Since the P-median is a problem of NP-hard, we have resolved the P-median with different approaches and strategies. The most recent result is a hybrid metaheuristic (Hybrid VNS/TABU), which consists in a smart combination of Variable Neighborhood Search (VNS) and Tabu Search (TS) that collects the most important approximations [15]. In this strand, previous articles that helped to find solutions like those presented in VNS-TS, can be seen in [15, 16].

In this section, we have chosen data from OR-Library [17] and we have tested four algorithms that have given good results for geographical data: 1.-Simulated Annealing (SA), 2.-Variable Neighborhood Search (VNS), 3.-Bioinspired Variable Neighborhood Search (VNS-BIO) and a 4.-Tabu Search-VNS Hybrid (H-TS-VNS).

However, the partitioning method PAM (Partitioning Around Medoids), that is modeled like the P-median, attained similar results along with H-TS-VNS, but better results than the other metaheuristics for the OR-Library instances, in a favorable computing time.

Nevertheless, for bigger instances that represent real states in Mexico, H-TS-VNS has surpassed PAM in time and quality in all instances. We expose the behavior of these five different algorithms for the test matrices from OR-Library and real geographical data from Mexico.

Furthermore, we made an analysis with the goal of explaining the quality of the results obtained to conclude that PAM behaves with efficiency for the OR-Library instances yet is overcome by the hybrid when applied to real instances.

On the other hand, we have tested the two best algorithms (PAM and H-TS-VNS) with geographic data from Jalisco, Queretaro and Nuevo León. At this point, as we said before, their performance was different than the OR-Library tests. The algorithm that attains the best results is H-TS-VNS.

The following tables contains the tests for the OR-Library instances. The nomenclature is Cost=objective cost and T=time (seconds). The details of each one of the 40 instances can be seen in [17], they go from 100 to 900 objects and the values of P from 5 to 200.

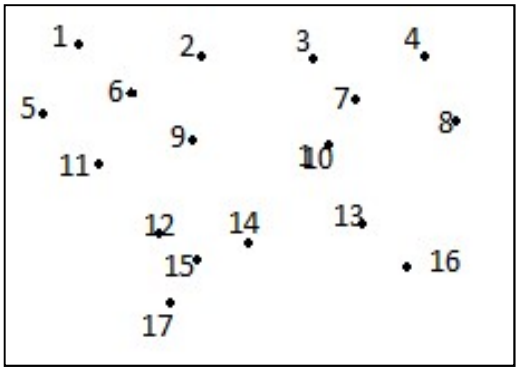


Fig. 1. Graph of 17 vertices.

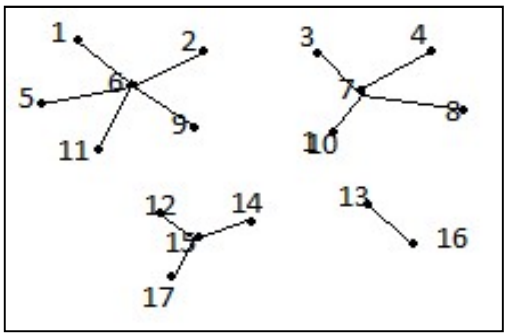


Fig. 2. Case of $n = 17$ with $p = 4$.

Table 3. Results for the OR-Library P-Median instances (Part 2).

Instance	PAM		H-VNS-TS	
	Cost	**T**	**Cost**	**T**
1	**5819**	0	**5819**	2.556
2	4105	0	**4093**	1.672
3	**4250**	0	**4250**	1.604
4	3046	1	3041	5.703
5	**1355**	1	1394	5.928
6	**7824**	0	**7824**	49.28
7	5645	1	**5631**	21.744
8	4457	2	4451	19.764
9	2753	8	2804	31.729
10	1263	14	1318	25.288
11	**7696**	0	**7696**	145.137
12	**6634**	1	**6634**	63.67
13	**4374**	20	4388	48.169

14	2974	56	3091	37.845
15	1738	82	1858	48.857
16	**8162**	1	**8162**	222.629
17	**6999**	2	**6999**	97.449
18	4811	67	4840	25.538
19	2859	296	2927	29.422
20	1805	600	1882	36.45
21	**9138**	0	**9138**	164.141
22	8669	4	8579	58.606
23	**4619**	160	4664	58.606
24	2965	938	3093	127.046
25	1844	1608	1937	132.722
26	**9917**	2	**9917**	389.316
27	**8307**	9	**8307**	68.365
28	4515	605	4551	35.594
29	3039	2101	3181	66.12
30	2009	2208	2119	105.318
31	**10086**	2	**10086**	479.083
32	**9301**	8	**9310**	109.158
33	4703	1495	4735	47.558
34	3026	4685	3168	119.309
35	**10400**	2	**10400**	413.429
36	**9934**	10	**9934**	141.098
37	5064	2092	5278	68.316
38	**11060**	8	**11060**	86.544
39	**9423**	13	**9423**	99.102
40	5138	5076	5214	76.852

The values in bold in both tables represent the tests that returned the best-known value (probably the global optimum) until today.

We can see that PAM achieves the best results, but its computing time increases as the problem size increases. The worst algorithm was SA.

Even though H-TS-VNS attains good results for the instances in Table 4 it's not a guarantee that it will be the same for the real geographical data of our interest. For this reason, we selected PAM and H-TS-VNS to test them with data from three states of Mexico: Jalisco, Querétaro and Nuevo León (3484, 814 and 2416). Our results are in Table 4.

For Table 4, we executed 9 instances, 3 for each map using 12, 24 and 48 P's (medians) for each. The instances 1, 2 and 3 are for the map of Querétaro that has 814 objects, instances 4, 5 and 6 are for Nuevo León with 2416 objects and 7, 8 and 9 are the instances for Jalisco, which has 3484 objects.

Table 4. Results for the geographical data.

Inst.	H-TS-VNS		PAM	
	Cost	**Time**	**Cost**	**Time**
1	**50.7595**	00:00:56	51.59754	00:02:03
2	**33.9236**	00:00:35	33.93228	00:15:33
3	23.7555	00:00:36	**23.338**	00:25:45
4	**210.9751**	00:08:05	211.2497	00:15:38
5	**139.5007**	00:04:21	140.1448	01:49:55
6	**94.5667**	00:02:15	**Didn't finish after 5 hours**	
7	**529.9198**	00:15:01	531.7125	00:24:28
8	**371.3131**	00:09:58	**Didn't finish after 5 hours**	
9	**243.5297**	00:05:51	**Didn't finish after 5 hours**	

We see in Table 4 that H-TS-VNS surpasses PAM in all the instances except for instance 3 (Querétaro with P = 48), however there's a big difference between the execution times. A peculiar aspect is that H-TS-VNS was run with the same input parameters for all the instances: nit = 1000, nit2 = 500, ip = 20 and tt = p/2 (see section 2.2.1.2 for more details) but its execution time considerably decreases.

This is because of our modified swap method where the candidate list is formed by the objects assigned to the selected medoid, therefore when the P is bigger the objects are more evenly distributed, generating smaller candidate lists for each medoid.

For example, for the map of Jalisco with P = 12 the algorithm finished in 15 minutes and for P = 48, in almost 6 minutes. Another aspect to consider is that the greedy method to generate the initial solution was only used for the map of Querétaro, because it required several minutes to generate the solution for the other maps, for this reason we used a randomly generated solution for Jalisco and Nuevo León but as we can see this did not negatively affect the quality of the final solution, as we still obtained better results than PAM.

These maps were generated with a custom Geographic Information Software developed in Java with the aid of GeoTools library, available for free in [1].

6. Conclusions

We observed that PAM surpassed the quality of all the other algorithms in the P-median instances from OR-Library. Even when we tried to give H-TS-VNS more time to find a solution, in many cases didn't manage to match the quality of PAM, not even do better than PAM in the instances where neither matched the best known result, however, as we saw in table 4 H-TS-VNS worked better than PAM, in quality and time, when working with geographical data, this is an issue that we need to examine in more detail to find the reasons for this behavior and make the necessary changes to achieve a consistent behavior with different kinds of data. For this we need to do more testing and debugging with our H-TS-VNS algorithm.

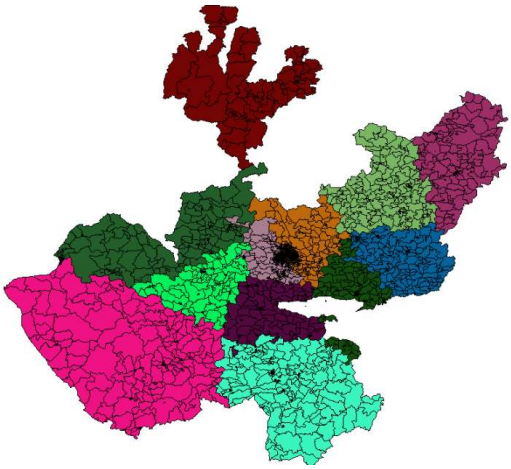


Fig. 3. Map of Jalisco with P = 12 returned by PAM. Cost: 531.7125.

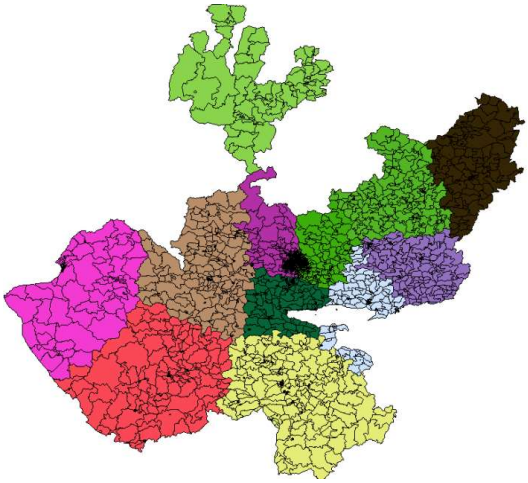


Fig. 4. Map of Jalisco with P = 12 returned by H-TS-VNS. Cost: 529.9198

The contribution of this article is to clarify that moreover than owning an equivalence the definition of zonas designed presented by Bação [19] and the classic definition of partitioning, it is also necessary to precise that the graphic solution generated by the model of the P-median answers to the fact that the P-median is a tool to cluster territories in problems of zones designs and in turn, the partitioning is a computational tool to solve DT problems.

Certainly, hierarchical clustering can be chosen instead of partitioning, situation that the authors will solve later. On the other hand, considering that the P-median model is integer-binary, for small instances, the commercial software solves the problem generating exact solutions.

However, for big problems, is insufficient the capacity of commercial software, then the inclusion of approximated methods is necessary to grant solutions close to the

optimum. In this point, with an example, a proposal of transformation of the P-median model from an integer binary model to a combinatorial one has been presented, the challenge is to generalize the transformation.

References

1. Shirabe, T.: A model of contiguity for spatial unit allocation. Geographical Analysis, vol. 37, no. 1, pp. 2–16 (2005) doi: 10.1111/j.1538-4632.2005.00605.x
2. Macmillan, W.: Optimization modelling in GIS framework: the problem of political redistricting. In: Fotheringham, S., Rogerson, P. (eds.) Spatial analysis and GIS, pp. 221–246 (1994)
3. Kalcsics, J., Nickel, S., Schröder, M.: Towards a unified territorial design Approach-Applications, algorithms and GIS integration. Top, vol. 13, no. 1, pp. 1–56 (2005) doi: 10.10 07/bf02578982
4. Hess, S. W., Samuels, S. A.: Experiences with a sales districting model: Criteria and implementation. Management Science, Institute for Operations Research and the Management Sciences, vol. 18, no. 4-part-ii., pp. P-41-P-54 (1971) doi: 10.1287/ mnsc.18.4.p41
5. Zoltners, A. A., Sinha, P.: Sales territory alignment: A review and model. Management Science, vol. 29, no. 11, pp. 1237–1256 (1983) doi: 10.1287/mnsc.29.11.1237
6. Beasley, J. E.: Welcome to OR-Library (2022) www.brunel.ac.uk/~mastjjb/jeb/info.html
7. Reese, J.: Solution methods for thep-median problem: An annotated bibliography. Networks, vol. 48, no. 3, pp. 125–142 (2006) doi: 10.1002/net.20128
8. Church, R. L.: COBRA: A new formulation of the classic p-median location problem. Annals of Operations Research, vol. 122, no. 1-4, pp. 103–120 (2003) doi: 10.1023/ a:1026142406234
9. Church, R. L.: BEAMR: An exact and approximate model for the p-median problem. Computers and Operations Research, vol. 35, no. 2, pp. 417–426 (2008) doi: 10.1016/ j.cor.2006.03.006
10. Drezner, Z.: Facility location: A survey of applications and methods. Springer (1995)
11. Daskin, M. S.: Network and discrete location: Models, algorithms, and applications, Second Edition, Wiley (2013) doi: 10.1002/9781118537015
12. ReVelle, C. S., Swain, R. W.: Central facilities location. Geographical analysis, vol. 2, no. 1, pp. 30–42 (2010) doi: 10.1111/j.1538-4632.1970.tb00142.x
13. Kariv, O., Hakimi, S. L.: An algorithmic approach to network location problems. II: The p-medians. SIAM Journal of Applied Mathematics, vol. 37, no. 3, pp. 539–560 (1979)
14. Romero-Montoya, M., Granillo, E., González-Velázquez, R., Bernabé-Loranca, M. B., Estrada-Analco, M.: A hybrid VNS/TABU search algorithm for solution the p-median problem. International Journal of Combinatorial Optimization Problems and Informatics, vol. 11, no. 2, pp. 67–74 (2020) ijcopi.org/ojs/article/view/137
15. Bernábe-Loranca, M. B., González-Velázquez, R., Granillo-Martinez, E., Romero-Montoya, M., Barrera-Cámara, R. A.: P-median problem: A real case application. Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 182–192 (2020) doi: 10.1007/978-3-030-49342-4_18
16. Bernábe-Loranca, M. B., Estrada-Analco, M., González-Velázquez, R., Martíne-Guzman, G., Ruiz-Vanoye: Location-allocation problem: A methodology with VNS metaheuristic. Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 1015–1024 (2019) doi: 10.1007/978-3-030-16660-1_99

17. Beasley, J. E: OR-Library (2014) people.brunel.ac.uk/~mastjjb/jeb/orlib/pmedinfo.html
18. GeoTools: GIS utilities library for Java (2014) www.geotools.org/
19. Bação, F., Lobo, V., Painho, M: Applying genetic algorithms to zone design. Soft Computing, vol. 9, no. 5, pp. 341–348 (2004) doi: 10.1007/s00500-004-0413-4

ISSN 1870-4069

Analysis of Public Databases of the Health Sector for Decision Making in Health Infrastructures through Artificial Intelligence

Agustín Grajales Castillo[1], Arieh Roldan Mercado Sesma[2], Virgilio Zúñiga Grajeda[3], Felipe Orozco Luna[4], Luis A. Medellín Serna[5], Raúl C. Baptista Rosas[2]

[1] Universidad de Guadalajara,
Centro Universitario de Tonalá,
Programa de Investigación Multidisciplinaria en Salud,
Mexico

[2] Universidad de Guadalajara,
Centro Universitario de Tonalá,
Departamento de Salud,
Mexico

[3] Universidad de Guadalajara,
Centro Universitario de Tonalá,
Departamento de Ciencias Computacionales,
Mexico

[4] Universidad de Guadalajara,
Centro de Análisis de Datos y Supercomputación,
Mexico

[5] Universidad de Guadalajara,
Centro Universitario de Ciencias Exactas e Ingeniería
Mexico

agustin.grajales3462@alumnos.udg.mx, forozco@cads.udg.mx,
{arieh.mercado, virgilio.zuñiga, luis.medellin,
raul.baptista}@academicos.udg.mx

Abstract: One of the priorities that requires urgent responses in Mexico is health. Currently, decisions to build hospitals and health centers are general and centralized. However, they should be based on existing evidence in the information available from public health databases such as distances between municipalities, number and levels of hospitals, poverty and mortality levels, to name a few, and not only on political decisions. This research focused on analyzing with artificial intelligence methods, (decision trees, random forests and bagging technique) of three disaggregated five-year periods, and obtaining a new level of knowledge based on the interpretation of data, for effective decision-making in health infrastructures in the state of Jalisco, Mexico. Decision trees and random forests give us high precision to disperse health infrastructures favoring more inhabitants. The statistical analysis carried out guides a change in the allocation of health systems. The results of this work demonstrate the need to direct/distribute health centers based on these findings.

Keywords: Artificial intelligence; machine learning; big data; decision making; health infrastructures.

1 Introduction

The national health situation in Mexico is extremely delicate. The hospital infrastructure is overwhelmed by the demand for services. While access to health services is restricted to the eligible population: 45% of households in Mexico, the rest, the population with economic resources or informal capacity, are served in private health schemes (out-of-pocket expenses), or they are not taken care of [1].

Geography also helps to complicate the problem, in the main cities, where the specialized infrastructure is located, it is not very accessible for a large part of the population, thus, to the complicated problem of accessing specialized services, distance is added, which represents critical and valuable time.

This directly affects the greater need for care in patients of different age groups and high degrees of specialization that currently do not exist, plus a budgetary burden that announced cuts to the health system [1, 2].

Currently, the official formats for the creation of health infrastructures go through the reading of procedures and filling in imprecise or biased formats [3, 4], which contributes to creating hospitals of a certain level in regions where it is not required or far from a large part of the population.

In the world, Korea and Japan have the highest rate of beds per 1,000 inhabitants, above 12. Followed by some economies in Europe with more than 4, United States 2.8; and the best positioned in Latin America is: Chile (2.01) Colombia (1.69) and Costa Rica (1.15) while in Mexico it is 0.9 [5, 6]. At the national level, Mexico City is the one with the highest availability of beds (2.4) followed by Nuevo León (1.3) and Jalisco (1.2).

Access to healthcare is a requirement for human well-being that is constrained, in part, by the allocation of healthcare resources relative to the geographically dispersed human population [7]. Furthermore, inequities in access to health care contribute to persisting disparities in health care outcomes [8].

It is well known that not all the population has access to a car, so they must reach a health center by public transport, which leads to another factor added to the problem of accessibility: a greater impact on health per se; due to circumstances attributable to traffic given that public transport infrastructure affects road traffic volumes and influences the choice of transport mode [9].

Although we are far from having universal access to health care, which would demand availability and accessibility of services for those who most need health services 10, the contribution of this research is relevant in the search to bring as much as possible, health infrastructures in the most remote or dispersed areas.

In China, Lan et al. published a study on the implications of building hospital infrastructure based on the needs and equipment that serve as a fundamental axis, that is, as a stage prior to the rest, with which the decision of the construction of the physical infrastructure per se, would be carried out if it is included in the hospital infrastructure promotion plans [11].

Table 1. Health establishments in Jalisco.

Type	First level	Second level	Third level	Others[1]
Quantity	1,492	245	16	81

These decisions, as in Mexico, are usually centralized, which also leaves a dilemma: build from the needs but that are not in the budget or, build infrastructures and do not have the budget to equip them. Concepts such as equity and efficiency are sometimes not compatible in health settings. The use of *artificial intelligence*, enabled the use of large amounts of data [12, 13, 14] in all sectors. In medicine, it marks the beginning of a strong impact in three aspects:

1 in clinics, the rapid interpretation of images with high degrees of accuracy;
2 in healthcare systems [15], improving workflow and the potential to significantly reduce medical (human) errors; and;
3 in patients, to enable them in the self-monitoring process for health promotion [16].

This research evaluates the information available in open access databases in the health sector, with artificial intelligence methods to generate predictive models and solve health infrastructure problems, with better evidence-based decision-making and trigger key access points to the health infrastructure, and provide coverage to the currently dispersed or unprotected population of these services.

2 Health Infrastructure in Jalisco

Jalisco is one of the most important economies in Mexico. Its capital: Guadalajara is the third most populated after Mexico City and Mexico state, respectively; is considered the city with the greatest potential for attracting investment in Mexico and second in economic potential in North America and together with five other entities share the largest *GDP* (Internal Product Gross).

Jalisco has 125 municipalities, an educational platform that integrates more than 140 higher level institutions, among which the nearly 50 campuses of the main 30 Universities stand out; the second oldest and largest University in the nation; 6 private universities of national dimension and strong international projection; 2 Advanced Research Centers, as well as one of the largest and most integrated networks of technological education institutions in Mexico [17].

Guadalajara together with nine other cities around it form the metropolitan area that cluster the most important *infrastructure of health* from public to private hospitals (supplemental material S1).

Due to the population of the eastern state of Jalisco must move to other places where the infrastructure and specialized services are found (second and third level) resulting in longer travel times, man-hours of no labor production, and above all, demand for specialized personnel.

[1] Warehouses, laboratories, vaccination centers, administrative offices

Table 2. Third level health establishments in Jalisco.

Municipalities	Guadalajara	Zapopan	Zapotlanejo	Tlajomulco	Tala	San Miguel	Ocotlán	Total
Third level	6	5	1	1	1	1	1	16

Table 3. Second level health establishments in Jalisco.

Municipalities	Guadalajara	Zapopan	Arandas	Zapotlán	Puerto Vallarta	Tepatitlán	Tonalá	Tlaquepaque	Ocotlán	Tlajomulco	Total
Second level	88	26	9	9	9	7	5	5	5	4	167

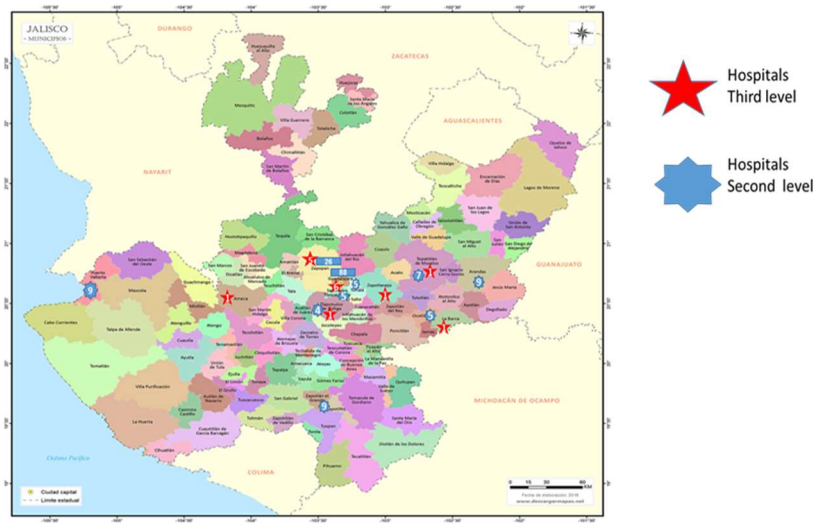


Fig 1. Municipalities of Jalisco. Adapted from: municipalities of Jalisco map with names - Google Search.

The state of Jalisco has 1,834 Health establishments, called CLUES (Unique Code of Health Establishments, IIEG) [18], classified as presented in table 1. While 511 are classified as Private Medical Services.

Tables 2 and 3 show the distribution of health infrastructures in the state; 7 municipalities concentrate 100% of level 3 and 10 municipalities 68% of level 2 hospitals. These levels of care are the basis of the National Health System. The concentration of these hospitals is in the metropolitan area of Guadalajara (figure 1).

From this perspective, it is essential to propose more comprehensive solutions, which go beyond installing orientation modules, accessibility and updated quasi-electronic files, among other things. These decisions must be supported by evidence considering the main variables: geography, distance between municipalities, number of hospitals, mortality levels and poverty levels, among others, and not only in political decisions.

However, the secretaries or government agents who decide on the growth of the infrastructure use few of these variables to resolve the creation of second and third level health centers or hospitals [27, 28], due to the large amount of data that needs to be processed. They simply do not analyze data.

Fig 2. Workflow diagram.

Table 4. Selection and description of variables.

Variable	Description
Poblacion	population in the five-year period
Pobreza	population in poverty in the five-year period
Defunciones	deaths in the five-year period
TasaMortalidad	deaths / population in the five-year period
PoblHosp	population / sumHosp1,2 in the five-year period
DefvsPob	deaths / poverty in the five-year period
DensidadxHabitante	population / surface km2
Medicos100kJal	260 doctors / 100, 000 inhabitants *[2]
NumHosp	number of 2nd and 3rd level hospitals
DistMun	90 km intermunicipal distance
NumMun	municipality number
Zone	zone
Mun	municipality name

This research focuses on how *AI* and *ML* [19] technological tools, can contribute to the analysis, design and modeling of data, in the municipalities of the state of Jalisco, to diagnose optimal solutions generation of health infrastructures; with the available information contained in free access databases related to the health sector.

The model intends to detonate key points, for the solution of new access points to the health infrastructure and cover the population currently dispersed or unprotected from these services.

3 Methods

3.1 Software Used

The programming language R version 4.1.3[3] [20] and the RStudio platform version 4.1.1[4] were used [21], which is a language with a statistical approach to data science.

3.2 Workflow Diagram

The workflow diagram is presented in figure 2.

[2] Taken from [27]

[3] https://cran.r-project.org/

[4] https://www.rstudio.com/

Table 5. Some variables do not provide weighty information to the model.

sex	age	scolarship	occupation	birthday	day of certificacion
necropsy	marital status	successor	preagnancy	birth month	month of certification

3.3 Data Transformation

Data mining [22] through data filtering, analysis and visualizations was applied in the work files to transform the information and make it readable by the software (figure 2).

Quantitative variables of population, poverty, deaths, mortality rate, population by hospital, deaths by population, per capita density, doctors per 100,000 inhabitants in Jalisco, number of hospitals, distances between municipalities, number of municipalities, zone and name of municipalities, of three disaggregated five-year periods (2010, 2015, 2020 INEGI) [23, 24, 25] were considered; applying statistics and machine learning classification algorithms: decision trees, random forests and bagging technique, to contrast results.

Although there are various techniques for classification, *decision trees* and *random forest* due to their short training time and results with high values of precision in classification are obtained. While other methods such as *support vector machines (SVM)* technique to classify has in the background the idea of finding the best hyperplane by which to separate the data. They work quite well for text classification, for example.

3.4 Variables Selection and Feature Extraction

Were include (13/50) variables [24, 25] (table 4), that they would contribute to the development of the model; and exclude others, with this, eliminate the "malediction of multidimensionality" [26] (reduction of redundant, spurious variables or those that do not provide weighty information to the model, see table 5). In this way, the relationships between the variables are better understood.

3.5 Processing

Were grouped municipalities in the variables NumMun (number of municipality), Zone, Mun (name of municipality) that were not more than 90 km apart (DistMun); (the above as a guideline that each access to a health infrastructure be 90 minutes away, considering a speed of 60 km/hour) and the number of second and third level hospitals (NumHosp).

The rest of variables were categorized into low, medium, high levels; to identify the different possibilities of variation. *Decision trees* and *random forests* are generated, and also the *bagging* technique, data partition and prediction (supplemental material S2), of the three five-year periods disaggregated with different variables and the statistical results are reflected.

Table 6. Results with the three algorithms of the three five-year periods.

Algorithm	Variables	Period	Sensitivity %	Specificity %	Accuracy %
Decision trees	10[5]	five-year period 2010-2014	100	100	100
Random forest	8[6]	five-year period 2015-2019	100	75	88
Decision trees	9[7]	five-year period 2020	100	100	100
Bagging	9[5]	five-year period 2010-2014	100	75	88

4 Results

The results are presented with the three algorithms of the three five-year periods with 10, 8 and 9 variables respectively, see table 6. We used the metrics of Sensitivity, Specificity (correct/incorrect classification) and balanced Accuracy.

For the five-year period 2010-2014, with 10 variables, the best result is presented in the *decision trees* at 100%: Zapotiltic (table 7 and map 1 supplemental material S3) would benefit 23 municipalities in 4 zones, covering a population of 439,904 inhabitants.

For the five-year period 2015-2019, with 8 variables, the best result is presented in *random forest* at 88%: San Miguel El Alto (table 8 and map 2 supplemental material S3) would benefit 19 municipalities in 3 zones, covering a population of 787,740 inhabitants. For the five-year period 2020, with 9 variables, the best result is presented in *decision trees* at 100%: La Barca (table 9 and map 3 supplemental material S3) would benefit 15 municipalities in 3 zones, covering a population of 960,253 inhabitants.

With ten variables, the models behave more stable, except for *bagging* (multiclassifier), which is stricter. With eight variables, the best results were in *random forests* at 88% and with nine variables the best results were in *decision trees* at 100% while *bagging* technique reached at 88% as highest result.

Population, poverty and deaths are recorded in all the models as they are the center of the analysis. Categorizing variables offers an important contribution (Figures A1 and A3, supplemental material S4). Mortality rate and deaths by population had the highest weight of in the models.

There is homogeneity in the variables. The different percentages of *Accuracy* in the results concentrate guide to maintain an objective of 80% in the different algorithms to consider the model stable, robust and with the possibility of improvement by incorporating and categorizing more weight variables.

[5] Number of municipalities, zone, name of municipalities, distances between municipalities, poverty, deaths, population, mortality rate, population by hospital, deaths by population

[6] Number of municipalities, zone, name of municipalities, distances between Municipalities, poverty, deaths, population, mortality rate

[7] Number of municipalities, zone, name of municipalities, distances between municipalities, poverty, deaths, population, per capita density, doctors per 100k inhabitants in Jalisco

A classification system is considered useful when it has a higher accuracy rate in the majority class (positives vs. negatives). The current processes to detonate *health infrastructures* in Mexico do not consider statistical methods, so this percentage is considered highly significant for decision making.

Supplementary material S3 shows the tables with the zones, municipalities, distances between municipalities, the coverage population of the municipalities, and corresponding coverage maps after the analysis. In addition, the map of the state of Jalisco with the three models after the analysis.

5 Conclusions

In the explored databases, there is a lot of useful information for the purposes of this research, there are variables that highlight the importance of considering the distances between the municipalities, the number of hospitals, deaths and poverty levels; it is well known that the higher the level of poverty, the greater the marginalization and, consequently, deaths. It is clarified that there are some other data that do not represent the leverage for the development of the model, as explained in tables 4 and 5; however, these latest data could be useful for health decision makers to better equip hospitals according to the type of disease, for example.

The national health situation in Mexico requires concrete and reliable responses in health systems. People's lives are involved. They must be based on a base of concrete, logical information and, where required, forecast and/or prevention. It is in these scenarios where *decision-making* 27 has become an “art”, and they cannot occur without first comprehensively analyzing the information available in the Health databases. In the work carried out to date, the advances presented suggest that, with an optimal combination of variables, and statistical and algorithmic *data analysis*, the resulting information is much more objective and precise.

It is evident that forecasting an infrastructure and not installing it (which would have saved lives) is not the same as installing it when it is not efficient in terms of serving the inhabitants, distances between municipalities, among other metrics. The results of the work demonstrate the need to direct/distribute the centers based on these findings.

As seen in the tables and maps generated in supplemental material S4, it is essential to cover the eastern areas of the state of Jalisco, since the actual distribution is concentrated in the metropolitan area, leaving these municipalities out and contributing to social inequality and economic backwardness. With this, the number of available beds in the state would rise, since, it is currently 1.2 [6].

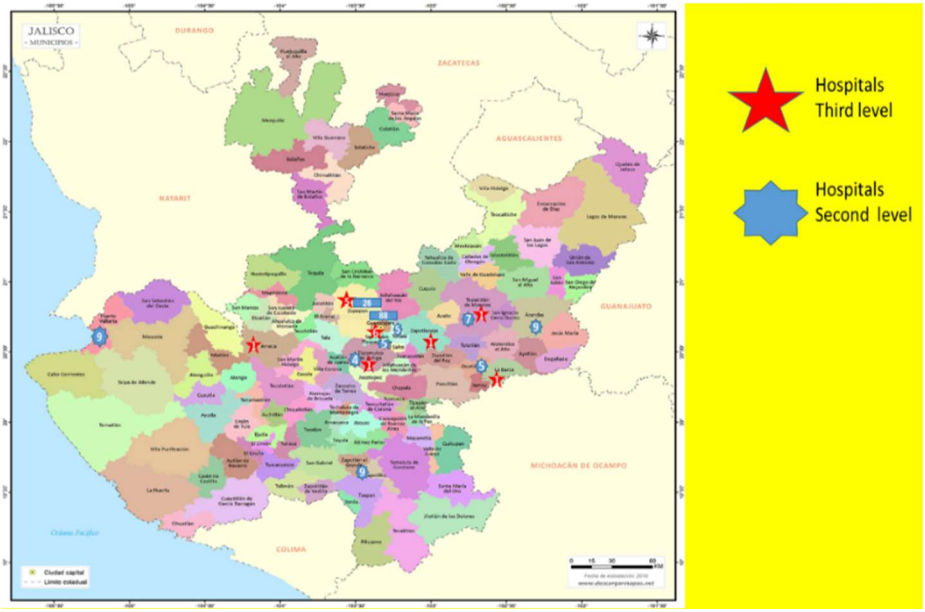
In this multidisciplinary line of research related to the combination of *machine learning* models and *data mining* [22, 29, 30, 31], it is aimed at solving problems and decisions in *health infrastructures*, with a statistical methodology assisted by the *AI*.

The social and scientific value that the research represents is important, since it is aimed at improving the living conditions and well-being of the population of the state of Jalisco; while producing knowledge that opens opportunities for *decision making* and problem solving in health service infrastructures.

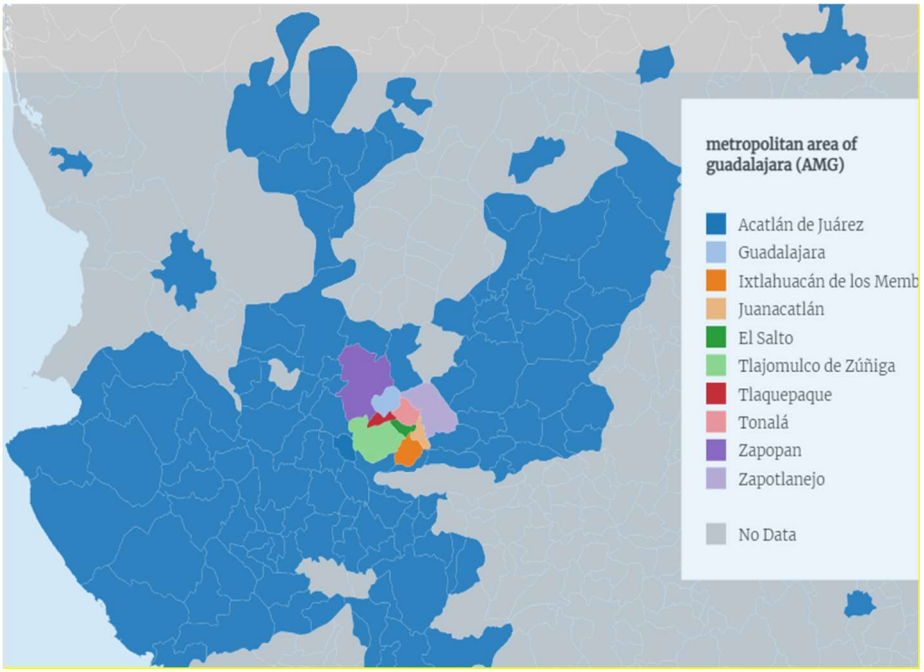
Supplementary materials are presented in the following sections: S1, S2, S3, S4.

Supplementary Material S1

(1) Distribution map of health infrastructures in Jalisco, second and third level and (2) Metropolitan area of Guadalajara (MAG) map.



(2) Figure adapted of: municipalities of Jalisco map with names - Google Search.



Supplementary Material S2

1) Pseudocode for grouping of municipalities and variable categorization.

```
temp <- dataset_distancias[dataset_distancias[,i] <= 90,]
colnames(temp)[i] <- "Distancia"
temp <- temp[,c(3,i)]
temp <- merge(temp,dataset, by=c("Municipio","Municipio"))

temp$factor_Pob220 <- discretize(temp$Poblacion2020,breaks=2,labels=c("Bajo","Alto"))
temp[,3] <- NULL
temp$factor_Pobre220 <- discretize(temp$Personas.en.situacion.de.pobreza2020..,breaks=2,labels=c("Bajo","Alto"))
temp[,3] <- NULL
temp$factor_Def220 <- discretize(temp$Defunciones2020..,breaks=2,labels=c("Bajo","Alto"))
temp[,3] <- NULL
temp$factor_TasaM <- discretize(temp$TASAS.DE.MORTALIDAD.x.100.000,breaks=2,labels=c("Bajo","Alto"))
temp[,3] <- NULL
temp$factor_PobHosp220 <- discretize(temp$PobHospQ2020,breaks=2,labels=c("Bajo","Alto"))
temp[,5] <- NULL
temp$factor_DefvsPobQ2020 <- discretize(temp$DefvsPobQ2020,breaks=2,labels=c("Bajo","Alto"))
temp[,5] <- NULL
```

2) Pseudocode for the corresponding algorithms (2.1 *decision trees*, 2.2 *random forest* and 2.3 *bagging*).

2.1)

```
# trees decision

    data <- data %>%
        mutate_at("random_nuevHosp_f", factor)

 split = sample.split(data$random_nuevHosp_f, SplitRatio = 2/3)

data_entrenamiento = subset(data, split == TRUE)
data_prueba = subset(data, split == FALSE)

arbol <- rpart(formula = random_nuevHosp_f ~ .,
        data = data_entrenamiento,
                    minsplit = 2,
                  method = "class")
```

2.2)

```
# random forest

    data <- data %>%
        mutate_at("random_nuevHosp_f", factor)

    split = sample.split(data$random_nuevHosp_f, SplitRatio = 2/3)

data_entrenamiento = subset(data, split == TRUE)
data_prueba = subset(data, split == FALSE)

arbol <- randomForest(formula = random_nuevHosp_f ~ .,
        data = data_entrenamiento)

# prediction and confusion matrix

pred_arbol <- predict(arbol, newdata = data_prueba, type = 'class')
data_prueba <- cbind(data_prueba,pred_arbol)

cm <- confusionMatrix(pred_arbol, data_prueba[["random_nuevHosp_f"]])
tocsv <- data.frame(cbind(t(cm$overall),t(cm$byClass)))
tocsv_percent <- tocsv
```

2.3)

```
# bagging technique

data <- data %>%
    mutate_at("random_nuevHosp_f", factor)

split = sample.split(data$random_nuevHosp_f, SplitRatio = 2/3)

data_entrenamiento = subset(data, split == TRUE)
data_prueba = subset(data, split == FALSE)

arbol <- bagging(formula = random_nuevHosp_f ~ .,
data = data_entrenamiento)
```

Supplementary Material S3

Tables 1-3 and the corresponding coverage maps (1-3) after the analysis and (4) general map after the analysis.

Table 1. Five-year period 2010-2014 after analysis.

NumMun	Zone	Municipalities	Zapotiltic (interdistances)	Population 2020
14004	Lagunas	Amacueca	58	5,770
14014	Lagunas	Atoyac	53	8,730
14023	Sur	Ciudad Guzmán	14	116,094
14057	Sureste	La Manzanilla De La Paz	85	4,118
14059	Sureste	Mazamitla	70	14,629
14065	Sur	Pihuamo	58	11,595
14069	Sureste	Quitupan	90	7,770
14079	Sur	San Sebastián Del Sur	33	17,932
14082	Lagunas	Sayula	43	37,902
14085	Sur	Tamazula De Gordiano	22	36,253
14086	Lagunas	Tapalpa	86	20,770
14087	Sur	Tecalitlán	32	17,365
14089	Lagunas	Techaluta De Montenegro	59	4,091
14092	Lagunas	Teocuitatlán De Corona	79	10,708
14099	Sur	Tolimán	75	11,928
14102	Sierra de Amula	Tonaya	80	5,989
14103	Sur	Tonila	37	7,600
14106	Sierra de Amula	Tuxcacuesco	80	5,508
14108	Sur	Tuxpan	12	38,573
14112	Sureste	Valle De Juárez	79	6,180
14113	Sur	San Gabriel	65	15,933
14119	Lagunas	Zacoalco De Torres	80	26,965
14122	Sur	Zapotitlán De Vadillo	72	7,501
			23	439,904

Map 1. Coverage map five-year period 2010-2014 after analysis

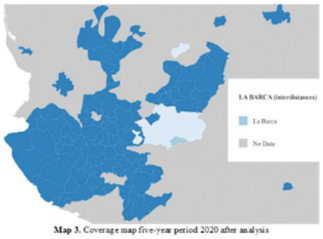
Table 2. Five-year period 2015-2019 after analysis.

NumMun	Zone	Municipalities	Sn Miguel El Alto (interdistancias)	Population 2020
14001	Altos Sur	Acatic	72	20,644
14008	Altos Sur	Arandas	57	79,688
14013	Ciénega	Atotonilco El Alto	69	56,994
14035	Altos Norte	Encarnación De Díaz	68	48,673
14046	Altos Sur	Jalostotitlán	19	30,917
14048	Altos Sur	Jesús María	78	18,227
14053	Altos Norte	Lagos De Moreno	78	177,818
14060	Altos Sur	Mexticacán	63	5,332
14072	Altos Norte	San Diego De Alejandría	46	7,645
14073	Altos Norte	San Juan De Los Lagos	35	70,418
14074	Altos Sur	San Julián	25	15,284
14091	Altos Norte	Teocaltiche	63	34,545
14093	Altos Sur	Tepatitlán De Morelos	50	133,350
14105	Ciénega	Tototlán	89	20,846
14109	Altos Norte	Unión De San Antonio	60	19,914
14111	Altos Sur	Valle De Guadalupe	27	6,658
14117	Altos Sur	Cañadas De Obregón	40	4,408
14118	Altos Sur	Yahualica De González Gallo	84	19,169
14125	Altos Sur	San Ignacio Cerro Gordo	54	17,[illegible]
			19	787,740

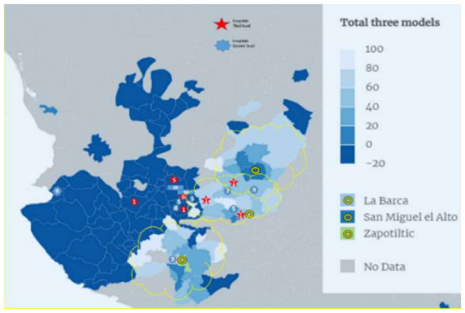
Map 2. Coverage map five-year period 2015-2019 after analysis

Table 3. Five-year period 2020 after analysis.

NumMun	Zone	Municipalities	La Barca (interdistances)	Population 2020
14008	Altos Sur	Arandas	69	79,688
14013	Ciénega	Atotonilco El Alto	34	56,994
14016	Ciénega	Ayotlán	49	42,119
14033	Ciénega	Degollado	70	23,954
14044	Centro	Ixtlahuacán De Los Membrillos	88	65,732
14047	Ciénega	Jamay	18	23,411
14048	Altos Sur	Jesús María	67	18,227
14063	Ciénega	Ocotlán	30	103,173
14066	Ciénega	Poncitlán	47	52,955
14070	Centro	El Salto	87	239,313
14093	Altos Sur	Tepatitlán De Morelos	79	133,350
14105	Ciénega	Tototlán	44	20,846
14123	Ciénega	Zapotlán Del Rey	48	17,697
14124	Centro	Zapotlanejo	76	65,584
14125	Altos Sur	San Ignacio Cerro Gordo	66	17,210
			15	960,253



Map 3. Coverage map five-year period 2020 after analysis



Map 4. Coverage map of the three models generated after analysis and their area of influence; second and third level hospital are represented.

SupplementaryMaterial S4

Variables categorization and contribution in generated models.

Figure A1

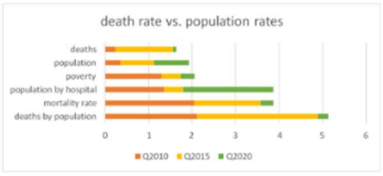


Figure A2

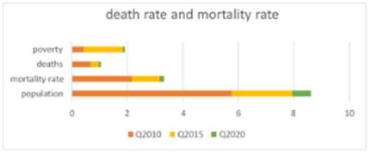
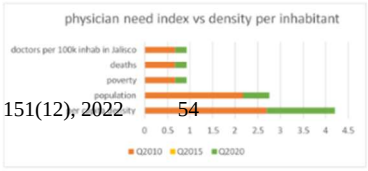


Figure A3



References

1. Jiménez, M.: Sistema de salud pública, un problema más para los mexicanos. Forbes México (2016) https://www.forbes.com.mx/sistema-de-salud-publica-un-problema-mas-para-los- mexicanos/
2. García, A.: Los retos para mejorar el Sistema de salud pública en México. El Economista (2019) https://www.eleconomista.com.mx/estados/Los-retos-para-mejorar-el-sistema-de-salud-publica-en-Mexico-20190105-0001.html
3. Secretaría de Salud: Plan maestro de infraestructura física en salud. Dirección General de Planeación y Desarrollo en Salud (2023) https://www.gob.mx/salud/acciones-y-programas/plan-maestro-de-infraestructura-fisica-en-salud
4. Secretaría de Salud: Certificado de necesidad. Dirección General de Planeación y Desarrollo en Salud (2022) https://www.gob.mx/salud/acciones-y-programas/certificado-de-necesidad?state=published
5. Statista: Hospital bed density select countries 2020. Health Proffesionals and Hospitals (2020) https://www.statista.com/statistics/283273/oecd-countries-hospital-bed-density/
6. Secretaria de Salud: Sistema de Información de la Secretaría de Salud (2018) http://sinaiscap.salud.gob.mx:8080/DGIS/
7. Weiss, D. J., Nelson, A., Vargas-Ruiz, C. A., Glicoric, K., Bavadekar, S., Grabilovich, E., Bertozzi-Villa, A., Roizer, J., Gibson, H. S., Shekel, T., Kamath, C., Lieber, A., Schulman, K., Shao, Y., Qarkaxhija, V., Nandi, A. K., Keddie, S. H., Rumisha, S., Amratia, P., Arambepola, R., Chestnutt, E. G.: Global maps of travel time to healthcare facilities. Nature Medicine, vol. 26, pp. 1835–1838 (2020) doi: 10.1038/s41591-020-1059-1
8. Guo, J., Hernandez, I., Dickson, S., Tang, S., Essien, U. R., Mair, C., Berenbrok, L. A.: Income disparities in driving distance to health care infrastructure in the United States: A geographic information systems analysis. BMC Research Notes, vol. 15, no. 225 (2022) doi: 10.1186/s13104-022-06117-w
9. Tétreault, L. F., Eluru, N., Hatzopoulou, M., Morency, P., Plante, C., Morency, C., Reynaud, F., Shekarrizfard, M., Shamsunnahar, Y., Faghih-Imani, A., Drouin, L., Pelletier, A., Goudreau, S., Tessier, F., Gauvin, L., Smargiassi, A.: Estimating the health benefits of planned public transit investments in Montreal. Environmental Research, vol. 160, pp. 412–419 (2018) doi: 10.1016/j.envres.2017.10.025
10. Escamilla, V., Calhoun, L., Winston, J., Speizer, I. S.: The role of distance and quality on facility selection for maternal and child health services in urban Kenya. Journal of Urban Health, vol. 95, no. 1, pp. 1–12 (2018) doi: 10.1007/s11524-017-0212-8
11. Lan, T., Chen, T., Hu, Y., Yang, Y., Pan, J.: Governmental investments in hospital infrastructure among regions and its efficiency in China: An assessment of building construction. Frontiers in Public Health, vol. 9, no. 719839 (2021). doi: 10.3389/fpubh.2021.719839
12. García-Alsina, M.: Big data: Gestión y explotación de grandes volúmenes de datos. Editorial UOC (2017) http://digital.casalini.it/9788491167112
13. López-Murphy, J. J., Zarza, G.: La ingeniería del big data: Cómo trabajar con datos. Editorial UOC (2017)
14. Bonsón Ponte, E., Sierra Molina, G.: Intelligent Accounting: impact of Artificial Intelligence on accounting research and accounting information. In: Proceedings of the ITHURS, León, pp. 361–368 (1996)
15. Hoffman S.: The great promise of artificial intelligence for public health. Research and Innovation (2018) https://yfile.news.yorku.ca/2018/04/05/the-great-promise-of-artificial-intelligence-for-public-health/
16. Topol, E. J.: High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, vol. 25, no. 1, pp. 44–56 (2019) doi: 10.1038/s41591-018-0300-7

17. Instituto de información y estadística geográfica de Jalisco (IIEG) (2022) https://iieg.gob.mx/ns/
18. Sebastian, Y., Tiong, X. T., Raman, V., Fong, A., Then, P.: Advances in diabetes analytics from clinical and machine learning perspectives. International Journal of Design, Analysis and Tools for Integrated Circuits and Systems, vol. 6, no. 1, pp. 32–37 (2017)
19. CRAN: The comprehensive R archive network (2023) https://cran.r-project.org/
20. Posit: The open-source data science company (2023) https://www.rstudio.com/
21. Febles-Rodríguez, J. P., González-Pérez, A.: Aplicación de la minería de datos en la bioinformática. Acimed, vol. 10, no. 2, pp. 69–76 (2002)
22. Instituto nacional de estadística y geografía (INEGI) (2022) https://www.inegi.org.mx/datos/
23. Dirección general de información en salud: Defunciones: Datos abiertos. Secretaría de Salud (2022) http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_defunciones_gob mx.html
24. Consejo nacional de la política de desarrollo social (CONEVAL): Indicadores de pobreza municipal 2010-2015. Datos Abiertos (2019) https://www.coneval.org.mx/Paginas/principal.aspx
25. Beltrán-Pascual, M., Muñoz-Martínez, A., Muñoz-Alamillos, Á.: Redes bayesianas aplicadas a problemas de credit scoring. Una aplicación práctica, Cuadernos de Economía, vol. 37, no. 104, pp. 73–86 (2014) doi: 10.1016/j.cesjef.2013.07.001
26. Vargas, M.: Jalisco con solo 2.6 médicos por mil habitantes. Centro Universitario de Ciencias de la Salud (2022) https://www.cucs.udg.mx/noticias/archivo-de-noticias/jalisco-con-solo-26-m-dicos-por-mil-habitantes
27. Turban, E., Sharda, R., Delen, D., Aronson, J. E., Liang, T. P., King, D.: Decision support and business intelligence systems. Pearson College Division (2005)
28. Mikhaylov, S. J., Esteve, M., Campion, A.: Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. Philosophical Transactions of the Royal Society A Mathematical, Physical and Engineering Sciences, vol. 376, no. 2128, (2018) doi: 10.1098/rsta.2017.0357
29. Benítez, R., Escudero, G., Kanaan, S., Rodó, D. M.: Inteligencia artificial avanzada. Editorial UOC (2014)
30. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthcare Journal, vol. 6, no. 2, pp. 94–98 (2019) doi: 10.7861/futurehosp.6-2-94

ISSN 1870-4069

Non-bio-inspired Metaheuristics in Software Testing

Alfredo Delgado-Santiago[1], Angel J. Sánchez-García[1],
Marcela Quiroz-Castellanos[2]

[1] Universidad Veracruzana,
Facultad de Estadística e Informática,
Mexico

[2] Universidad Veracruzana,
Instituto de Investigaciones en Inteligencia Artificial,
Mexico

dltunasd@gmail.com, {angesanchez, mquiroz}@uv.mx

Abstract. The software testing phase usually consumes a large part of the development of software projects trying to get as many defects as possible in the final product. Different strategies have been approached to optimize this phase of the testing stage; such is the case of metaheuristics, algorithms with the ability to find high-quality solutions in a relatively short time. This research seeks to analyze the current status of the application of metaheuristics that assist in software testing phase activities, only the most representative non-bio-inspired algorithms (NBA) are surveyed, being Hill Climbing and Local Search the most used. The main activities of the software testing stage where NBA was implemented were test case generation, test data generation and test case prioritization (redundancy reduction). It was concluded that NBAs used on their own are only viable in some activities of the software testing phase. As future work, it is proposed to investigate the use of hybrid algorithms and approaches in software testing phase.

Keywords: Metaheuristic, software testing, optimization, systematic literature review.

1 Introduction

The need for software systems to be free of defects is increasing. To ensure the quality of the Software, a transcendental phase is the testing phase. In this revolution called Industry 4.0, Artificial Intelligence seeks to automate processes and provide products with autonomous decision-making, among other benefits.

In the software development process, there are studies on the use of metaheuristics at the software testing stage, however, most of these articles focus on bio-inspired algorithms (i.e., algorithms based on nature), and there are several other alternatives that could contribute to this field.

Table 1. Research questions.

Question	Motivation
RQ1.- What are the non-bio-inspired metaheuristics that have been reported at the software testing stage?	To identify the NBAs reported in the software testing phase.
RQ2.- What are the main activities of the testing stage where non-bio-inspired metaheuristics have been applied?	It is important to identify in which software testing activities the algorithms reported in RQ1 have been used, in order to analyze the contributions.
RQ3.- What are the advantages and disadvantages that have been found with the application of non-bio-inspired metaheuristics at the testing stage?	One of the objectives of this research is to describe the strengths and weaknesses of each approach to know in which activities they will be able to obtain better results.
RQ4.- What types of benchmark problems have been used to test non-bio-inspired metaheuristics?	To know the benchmarks used to test the algorithms found, to identify their characteristics and compare results with the proposals generated in future work.

A heuristic method is a tentative and plausible procedure whose purpose is to discover the solution to a particular problem. The heuristic (or simply heuristic) is a method that helps to discover the solution of a problem by making plausible but fallible guesses about what the best thing is to do next [1].

Building on the above, a metaheuristic is a high-level problem-independent algorithmic framework that provides a set of guidelines or strategies for developing heuristic optimization algorithms. Notable examples of metaheuristics include genetic/evolutionary algorithms, tabu search, simulated annealing, and ant colony optimization [2].

As the years go by, the systems developed in software projects become more complex, and, consequently, so do the tests. More and more alternatives are emerging to address the different optimization approaches in the development of software testing, such as the use of NBAs.

The use and research of these alternatives have shown that their use in the field of software testing could represent an optimization to this rigorous stage.

This paper is organized as follows: Section 2 describes background and related work. Section 3 details the method used to execute this Systematic Literature Review. In Section 4, the results obtained from this work are presented. Finally, Section 5 draws the main conclusions and proposes future work.

2 Background and Related Work

Artificial Intelligence is a discipline that has supported each of the phases of software development, such as requirements, design, coding, testing and maintenance.

In the testing phase, especially optimization algorithms have been used to generate test cases or identify defects. However, most of the strategies used are based on evolutionary algorithms or bio-inspired algorithms.

In a manual search of related work, a Systematic Review about Bio-inspired computation in software testing was found [3], which talks about the importance of

Table 2. Keywords and synonyms identified.

Concept	Synonyms and related terms
Metaheuristic	Metaheuristics, Meta-heuristic
Software engineering	-
Software testing	Testing
Benchmark	Benchmarks
Path algorithm	Trajectory algorithm
Local search	Explorative search
Neighborhood search	-
GRASP	Greedy Randomized Adaptive Search Procedure
Simulated annealing	-

software testing and how optimization processes are handled in the software testing stage and objective decision making. However, no Systematic Literature Review on the application of metaheuristics not based on evolutionary algorithms in Software Testing was identified.

Therefore, the purpose of this Systematic Literature Review (SLR) is to complement the identified systematic review with metaheuristics that are not bioinspired. With this, it will be possible to have a balance of both approaches, to describe advantages, disadvantages and possible combinations between them to improve the results obtained in the literature.

In addition, it is intended to identify software testing activities (such as test case generation, branch coverage, defect identification, among others), where different search optimization approaches have been used. Finally, it is expected to identify benchmarks of various software testing activities, in which different optimization and search approaches can be tested to compare future contributions.

3 Research Method

The method proposed by Kitchenham and Charters [4], which was proposed to carry out SLR Software Engineering area, was selected for this work. The planning phase is presented below.

3.1 Research Questions

This section shows the Research Questions (RQs) proposed for this research work. Research Questions are described in Table 1.

3.2 Search Strategy and Data Sources

This section shows the keywords and related terms that were used in the search strings. The decision to include the terms “Path Algorithms”, “Local Search”, “Neighborhood

Table 3. Data source.

Database	Website
IEEE Xplore	https://ieeexplore.ieee.org/Xplore/home.jsp
ACM	https://dl.acm.org/
Springer Link	https://link.springer.com/
Science Direct	https://www.sciencedirect.com/

Table 4. Inclusion criteria.

ID	Description
IC1	The study was published between 2017 and 2022.
IC2	Full access to the study.
IC3	The title or abstract of the study contains the search term 'Software testing' and its synonyms with another search term.
IC4	Reading the abstract, the study hints at answering at least one research question.

Search", "GRASP" and "Simulated Annealing" was made because they were considered important search terms in the context of this SLR.

The proposed search string from the key terms is described below.

("Software testing" OR Testing) AND ("Software engineering") AND ("trajectory algorithm" OR "Local search" OR "Explorative search" OR "Neighborhood search" OR "Greedy Randomized Adaptive Search Procedure" OR GRASP OR "Simulated Annealing" OR "Tabu search") AND (benchmark OR benchmarks).

Due to Science Direct operators limit (OR and AND), the string was adapted for use in this research source, so that the string was as similar as possible to the main string. The following search string was used in Science Direct.

"Software testing" AND "Software engineering" AND ("trajectory algorithm" OR "Local search" OR "Explorative search" OR "Neighborhood search" OR GRASP OR "Simulated annealing" OR "Tabu search").

Table 3 shows the databases used as source for this SLR.

3.3 Selection of Primary Studies

In this section, the criteria for the selection of primary studies are presented. The inclusion and exclusion criteria are presented in Table 4 and Table 5, respectively.

3.4 Selection Procedure

The selection procedure was made up of the following stages:

- Stage 1. Primary studies are filtered according to IC1 and IC2.
- Stage 2. The primary studies are removed according to EC1 and EC2.
- Stage 3. Primary studies are filtered according to IC3 and IC4.
- Stage 4. The primary studies are removed according to EC3.

Table 5. Exclusion Criteria.

ID	Description
EC1	Studies that are not written in the English language.
EC2	Studies that are outreach articles, posters, books, chapters, presentations, abstracts, or tutorials.
EC3	Duplicated studies.

Table 6. Application of inclusion and exclusion criteria by stage.

Database	First Results	Stage 1	Stage 2	Stage 3	Stage 4
ACM Digital Library	484	193	126	7	7
IEEE Xplore	38	18	18	4	4
SpringerLink	806	263	261	3	3
Science Direct	300	116	110	5	5
Total	**1, 628**	**590**	**515**	**19**	**19**

4 Results

The search process was carried out according to the SLR planning, executing the final search string in each of the selected sources. As it is shown in Table 6, the greatest reduction of studies occurred during Stage 3 of the selection process.

The list of references of the 19 primary studies selected for analysis can be found in [5]. The template for data extraction from each primary study can be found in [6].

Nineteen articles were obtained from the application of the search criteria; however, it was detected that only 9 of them contained information relevant to this research, since they answered at least one research question. Fig. 1 shows the proportion of the type of paper found (journal paper or conference paper).

It is worth mentioning that no dominant journal or conference was found. That is, each primary study belongs to a different Journal or conference. Fig. 2 shows the distribution of the years of publication of the selected studies, with 2018 being the most dominant year on this topic.

Next, the report of the results obtained by answering each research question is presented.

4.1 RQ1: What Were the Non-Bio-Inspired Metaheuristics that Have Been Reported at the Software Testing Stage?

Fig. 3 shows the frequency of the algorithms found in this investigation. Several types of NBA were identified: Hill climbing, Random search, Simulated annealing, Greedy Algorithm, LIPS (Linearly Independent Path-based Search), Neighborhood search and Ls-Sampling, being Search based approaches the most studied in the Software testing stage.

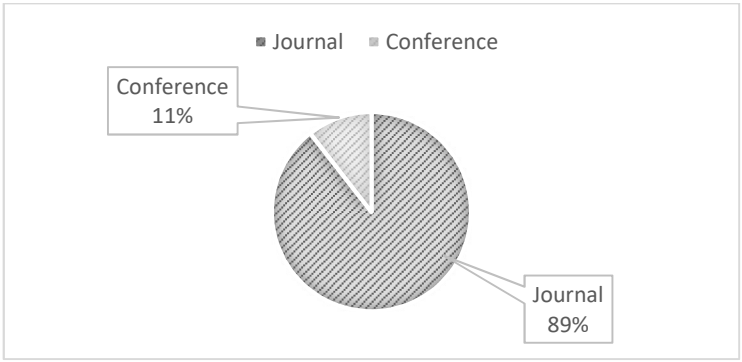


Fig. 1. Distribution by types of publication.

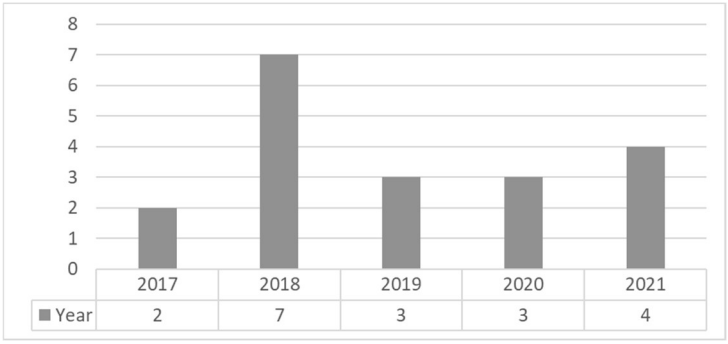


Fig. 2. Distribution by year in selected databases.

Hill climbing tends to be used for use case prioritization, as it is an optimization algorithm. Due to its variants, the Search-based approaches was used numerous times in the articles, such as Hill climbing, Random search, Neighborhood search and Ls-Sampling.

4.2 RQ2: What are the Main Activities of the Testing Stage where Non-Bio-Inspired Metaheuristics Have Been Applied?

The purpose of this research question is to identify the different activities of the testing stage where the Algorithms detected in **RQ1** were applied. Fig. 4 shows the most addressed testing phase activities.

Four activities were reported for the software testing phase where NBAs assist; most of the algorithms were used for test case prioritization; most of the algorithms detected are optimization algorithms. Hill Climbing [7, 8, 9], Greedy Algorithm [7, 9], Random Search [9], Simulated Annealing [8] and Neighborhood search [8] were used for this activity. Hill Climbing [10] and Simulated Annealing [10] algorithms were used for test data generation.

The findings of this research indicate that, compared with bio-inspired algorithms, NBAs non-bio-inspired algorithms did not perform well doing test cases generation. In this activity, the use of LIPS [11] was detected. According to the findings NBAs were

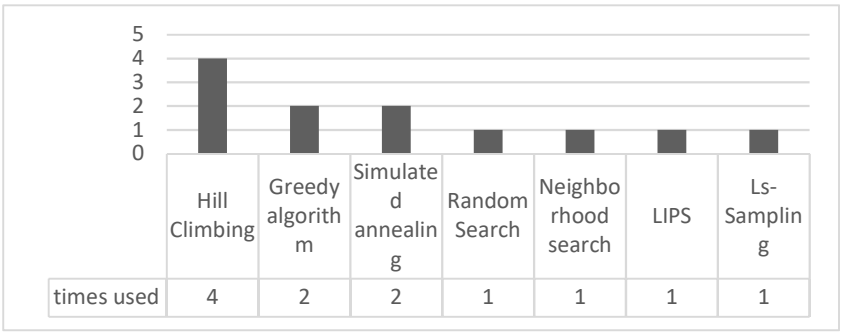


Fig. 3. Frequency of reported algorithms.

used to sort and group test cases into test suites. Only the use of Ls-Sampling [12] was reported for this activity of the software testing stage. LS-Sampling is a Search-based sampling approach. Search-based approaches showed the most versatility for performing software testing stage activities.

4.3 RQ3.- What are the Advantages and Disadvantages Found with the Application of Bio-Inspired Algorithms at the Testing Stage?

The algorithms reported in primary studies were mostly compared against bio-inspired algorithms; the results of these comparisons were that, commonly NBAs did not show a significant improvement compared to bio-inspired algorithms; however, due to their simplicity, i.e., the small number of lines of code required for their execution, these algorithms are optimal to apply to specific or reduced tasks.

In large systems, according to the findings, it is more advisable to use multi-objective approaches due to the number of functionalities they cover. Most of the reported algorithms were used to test case prioritization.

4.4 RQ4.- What Types of Benchmark Problems Have Been Used to Test Non-Bio-Inspired Metaheuristics?

Most of the NBAs were tested with specific problems, i.e., problems proposed by the authors to simulate a real-life scenario [9, 8, 11]. Triangle was used in two studies [9, 12]. One study was tested with the Corpus SF110 benchmark [12]. The results can be seen in Fig. 5.

5 Conclusions and Future Work

Industry 4.0 provides an automation of processes that help the manufacture of many essential products such as software. Nowadays, software systems are becoming more and more complex, therefore, testing those systems is becoming more and more complicated. So, the use of algorithms that help optimize software testing activities is a necessity.

Fig. 4. Activities of the testing phase.

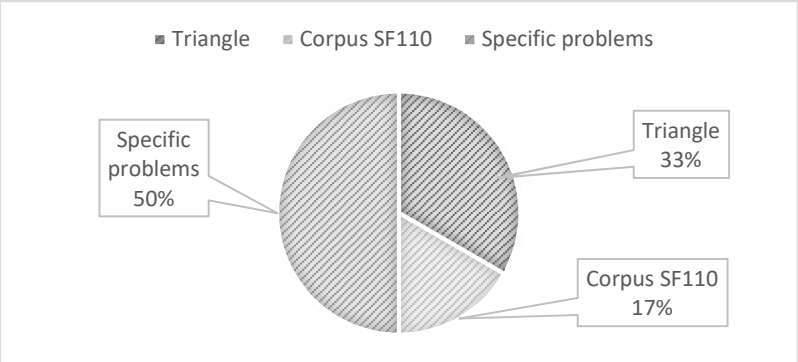


Fig. 5. Reported benchmark.

According to the information gathered in this SLR, non-bio-inspired metaheuristics mean a great improvement to the software testing process, however, bio-inspired algorithms are still superior in this aspect [12]. In carrying out this research, the use of hybrid algorithms was detected, showing a significant improvement (according to the benchmarks) in the efficiency when performing activities in the testing process.

It is proposed as future work the research of hybrid algorithms and approaches, since according to studies [13, 14] they represent a significant improvement over the use of individual approaches.

References

1. Feigenbaum, E. A, Feldman, J: Computers and thought. McGraw-Hill (1963)
2. Sörensen, K., Glover, F. W.: Metaheuristics. Encyclopedia of Operations Research and Management Science, pp. 960–970 (2013) doi: 10.1007/978-1-4419-1153-7_1167
3. Gómez-San-Gabriel, J. E., Sánchez-García, Á. J., Cortés-Verdín, K.: Cómputo bio-inspirado en la prueba de software: Una revisión sistemática de la literatura, vol. 149, no. 11, pp. 125-134 (2020)
4. Kitchenham, B. A., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University Joint Report (2007)

5. Appendix A: Primary studies references (2022) drive.google.com/file/d/1 PyDa70JBLd9jmNR_36YkFfdLziIXSJy/view?usp=sharing
6. Appendix B: Template for information analysis (2022) drive.google.com/file/ d/1ehs-f0SNZcDeZmc3PoBaFTY8ypM7FDk/view?usp=sharing
7. Lou, Y., Chen, J., Zhang, L., Hao, D.: A survey on regression test-case prioritization. Advances in Computers, Elsevier, pp. 1–46 (2019) doi: 10.1016/ bs.adcom.2018.10.001
8. Zamli, K.Z., Safieny, N., Din, F.: Hybrid test redundancy reduction strategy based on global neighborhood algorithm and simulated annealing. In: Proceedings of the 7th International Conference on Software and Computer Applications, Association for Computing Machinery, pp. 87–91 (2018) doi: 10.1145/3185089.3185146
9. Prado-Lima, J. A., Vergilio, S. R.: Search-based higher order mutation testing. In: Proceedings of the III Brazilian Symposium on Systematic and Automated Software Testing, Association for Computing Machinery, pp. 87–96 (2018) doi: 10.1145/3266003.3266013
10. Nosrati, M., Haghighi, H., Vahidi-Asl, M.: Using likely invariants for test data generation. Journal of Systems and Software, vol. 164, pp. 110549 (2020) doi: 10.1016/j.jss.2020.110549
11. Luo, C., Sun, B., Qiao, B., Chen, J., Zhang, H., Lin, J., Lin, Q., Zhang, D.: LS-sampling: an effective local search based sampling approach for achieving high t-wise coverage. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Association for Computing Machinery, pp. 1081–1092 (2021) doi: 10.1145/3468264.3468622
12. Panichella, A., Kifetew, F. M., Tonella, P.: A large scale empirical comparison of state-of-the-art search-based test case generators. Information and Software Technology, vol. 104, pp. 236–256 (2018) doi: 10.1016/j.infsof.2018.08.009
13. Lu, C., Zhong, J., Xue, Y., Feng, L., Zhang, J.: Ant colony system with sorting-based local search for coverage-based test case prioritization. In: IEEE Transactions on Reliability, vol. 69, no. 3, pp. 1004–1020 (2020) doi: 10.1109/ tr.2019.2930358
14. Monemi-Bidgoli, A., Haghighi, H.: Augmenting ant colony optimization with adaptive random testing to cover prime paths. Journal of Systems and Software, vol. 161, pp. 110495 (2020) doi: 10.1016/j.jss.2019.110495

ISSN 1870-4069

An Arduino FIST Evaluation for Fuzzy System Conversion from Matlab to Arduino

Jose Eleazar Peralta-Lopez, David Lazaro-Mata,
Jose Alfredo Padilla-Medina, Francisco Javier Perez-Pinal,
Alejandro Israel Barranco-Gutierrez

Tecnológico Nacional de México en Celaya,
Departamento de Electrónica,
Mexico

israel.barranco@itcelaya.edu.mx

Abstract. Today, the Arduino micro-controller is widely used because many engineers have contributed to the manufacture of different solutions that are documented and freely shared on the Internet. This digital circuit has pre-loaded typical functions that make it easy to use. Proof of this is the compiler built by the group of programmers from makeproto.com, they managed to build a translator from Matlab code to Arduino code which we put to the test. The purpose of this work is to quantitatively evaluate the implementation of a fuzzy system on Arduino from its Matlab design. The example presented is the implementation of a Mamdani-type fuzzy system to infer the tip given to a restaurant because is a very typical problem in fuzzy systems. The fuzzy system implementation on Arduino give a minimal numerical difference between systems but highlighted rapid fuzzy system implementation from Matlab designer.

Keywords: Arduino, fuzzy system, FIST, code translator.

1 Introduction

Currently, the electronics community has seen a fast development in the number and diversity of applications of Fuzzy Logic, ranging from consumer integrated circuit technology and industrial process control to decision support systems and financial trading.

These applications range from Navigation Control [1], ambiguity and noise diminishing from biometric security applications [2], stabilization of continuous-time the design of saturated sampled-data Parallel-Distributed Compensation controllers, composed by the estimation of the closed-loop domain of attraction [3], to voltage compensation in DC-DC converters [4]. So, the optimization of the implementation of fuzzy systems in embedded circuits has increased significance.

This leads us to evaluate their energy feeding, size, weight, speed of response and implementation time. There are also other implementations of fuzzy systems on raspberry pi, FPGA, typhon HIL or PCs [5, 6].

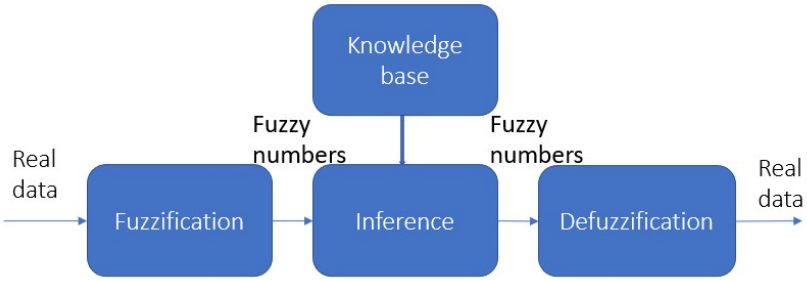


Fig 1. General structure of fuzzy systems.

Fig. 2. Arduino Pro-Mini.

Fuzzy systems have become relevant for being one of the artificial intelligences methods that best compete with neural networks due to the explainable of their architecture and the easy understanding of their working ranges [7].

Fuzzy systems mainly consist of four subsystems: fuzzification, knowledge base, inference engine, and defuzzification [8]. Fuzzification consists of converting sharp or real numbers into fuzzy numbers, in order to convert from the numerical world to the linguistic or ambiguous world.

The knowledge base contains the inference rules that the inference engine executes based on the input values. Finally, the defuzzification subsystem converts fuzzy numbers to sharp or real numbers as shows Fig. 1.

By other hand, Arduino free hardware boards are increasingly used for the development of control programs, especially academic level programs. But there are areas that have begun to be addressed. One of these areas is the implementation of fuzzy systems. Being able to implement this type of system on a board like the one mentioned brings advantages over other platforms due to its portability and low cost, which increases the advantage of using free hardware and software [9].

2 Materials and Method

In this work we basically use three tools: the Matlab fuzzy systems designer, the Matlab to Arduino compiler (Arduino FIST) and an Arduino Pro-Mini microcontroller. The group of programmers from http://www.makeproto.com managed to build a translator from Matlab code to Arduino code (an extension of "c" language) which we put to the test in this article [10].

The example presented is the implementation of a Mamdani-type fuzzy system to infer the tip given to a restaurant because is a very typical and basic problem in fuzzy systems.

2.1 Fuzzification

The definition on fuzzy system works as follow: Functions (1) to (5) show the description of different variables used. F represents qualification for "food", variable S represents "Service". Specifically, RF means "Rancid food", DF means "Delicious food", PS represents "Poor service", GS is for "Good service" and ES for "Excellent service":

$$RF = \begin{cases} 1 & 0 \leq f < 1, \\ \dfrac{3-f}{2} & 1 \leq f < 3, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

$$DF = \begin{cases} 0 & \text{otherwise}, \\ \dfrac{f-7}{2} & 7 \leq f < 9, \\ 1 & 9 \leq f < 10, \end{cases} \tag{2}$$

$$PS = \begin{cases} \dfrac{3-S}{3} & 0 \leq f < 3, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

$$GS = \begin{cases} \dfrac{S-2}{3} & 2 \leq f < 5, \\ \dfrac{8-S}{3} & 5 \leq f < 8, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

$$ES = \begin{cases} \dfrac{S-7}{3} & 7 \leq f < 10, \\ 0 & \text{otherwise}. \end{cases} \tag{5}$$

2.2 Knowledge Base and Inference Engine

In order to develop the knowledge base or rules of the system, the maximum and minimum fuzzy operators are used. These operators are equivalent to the logical operators OR and AND respectively.

The maximum operator compares two membership degrees of both inputs and assigns the maximum value to the defuzzifier, the minimum operator compares two membership degrees equally and the minimum value between both is assigned to the output. In the case of this fuzzy system, the rules implemented are the following:

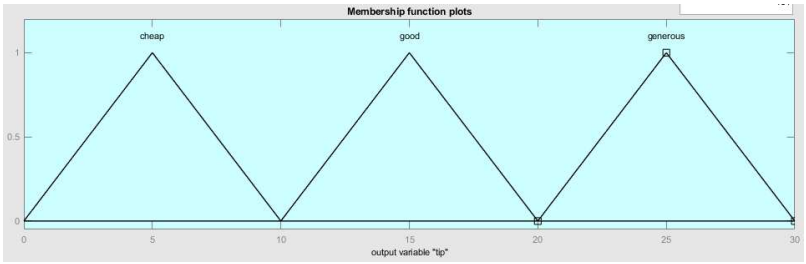


Fig. 3. Fuzzy output membership functions.

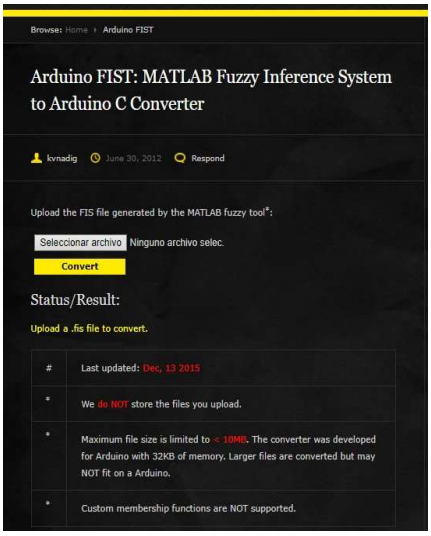


Fig. 4. Web application to convert the fuzzy system designed in Matlab to its Arduino version.

1 IF food is rancid or service is poor THEN tip is cheap.
2 IF service is good THEN tip is good.
3 IF food is delicious or service is excellent THEN tip is generous.

2.3 Defuzzification

In this case, the defuzzification is done by calculating the centroid of the figure resulting from applying different levels to each of the membership functions of the output, as can be seen in the equation (6) and Fig. 3, where the variable TIP is the output and the variables A_{tch}, A_{tgo} and A_{tge} represent the areas of the membership functions of the cheap tip, good tip and generous tip, respectively, contained from 0 to the degree of membership of each one.

The calculation of the area can be done with numerical methods or with approximations such as the one shown in [9]:

$$TIP = \frac{5 * A_{tch} + 15 * A_{tgo} + 25 * A_{tge}}{A_{tch} + A_{tgo} + A_{tge}}. \quad (6)$$

Table 1. Fuzzy systems outputs in Matlab and Arduino with its absolute difference.

Food	Service	Matlab	Arduino	Error
0	0	5	5	0
0	2	5	5	0
0	3	8.58	8.58	0
0	5	10	10	0
0	7	8.58	8.58	0
0	8	12.10	12.14	0.04
0	10	15	15	0
2	0	5	5	0
2	2	5	5	0
2	3	9.26	9.26	0
2	5	10.7	10.72	0.02
2	7	9.26	9.26	0
2	8	13.5	13.57	0.07
2	10	16.4	16.43	0.03
5	0	5	5	0
5	2	5	5	0
5	3	15	15	0
5	5	15	15	0
5	7	15	15	0
5	8	25	25	0
5	10	25	25	0
8	0	13.6	13.57	0.03
8	2	16.5	16.49	0.01
8	3	20.7	20.74	0.04
8	5	19.3	19.28	0.02
8	7	20.7	20.74	0.04
8	8	25	25	0
8	10	25	25	0
10	0	15	15	0
10	2	17.9	17.82	0.08
10	3	21.4	21.42	0.02
10	5	20	20	0
10	7	21.4	21.42	0.02
10	8	25	25	0
10	10	25	25	0

2.4 From Matlab code to Arduino

The fuzzy system is designed in the Fuzzy Logic Designer, it is saved as a .fis file, it is uploaded to the application and it gives us a .zip file that contains the c-arduino code and another .h file that together work to execute the system previously designed.

In the Fig 4 it shows the Web application to upload the fuzzy system designed in Matlab to get its Arduino version.

3 Results

Using the "Arduino FIST" web application, 304 lines of C-Arduino code were obtained and its h file header. Table 1 shows a set of different inputs to the fuzzy system implemented in Matlab and the other implemented in Arduino, as well as the difference between both. In order to compare the results and the quality of the implementation.

4 Discussion and Conclusions

It is concluded that the conversion of fuzzy systems from Matlab to Arduino using the "Arduino FIST" web tool produces results almost of the same quality as the MATLAB Fuzzy Toolbox.

With a maximal error of 0.08 for the particular case where the food was 10 and the service 2. It is convenient highlight the advantages of portability, low power consumption and low cost of Arduino and a Matlab to Arduino ("C") conversion speed of less than one second.

An improvement to this proposal could be the implementation of the accuracy improvement for numbers less than tenths of a unit. In other works, such as [11], this tool has been analyzed qualitatively, while in this document we present a quantitative analysis to observe the order of error that the fuzzy system implemented in Arduino gives us. For this case the maximum error in an estimate was 0.08.

Acknowledgments. The Authors want to thank CONACyT for the national scholarships granted and the encouragement of the Sistema Nacional de Investigadores. Also, we want to thank to the TecNM for the Doctorate, Master's and Bachelor's degree programs in Electronic Engineering in Celaya.

References

1. Mokhtari, K., Wagner, A. R.: Pedestrian collision avoidance for autonomous vehicles at unsignalized. Journal of Applied Technology and Innovation (2021)
2. Irshad, A., Usman, M., Chaudhry, S. A., Bashir, A. K., Jolfaei, A., Srivastava G.: Fuzzy-in-the-loop-driven low-cost and secure biometric user access to serve. IEEE Transactions on Reliability, vol. 70, no. 3, pp. 1014–1025 (2020) doi: 10.1109/TR.2020.3021794
3. Lopes, A., Guelton, K., Arcese, L., Leite, V.: Local sampled-data controller design for T-S fuzzy systems with saturated actuators. IEEE Control Systems Letters, vol. 5, no. 4, pp. 1169–1174 (2020) doi: 10.1109/LCSYS.2020.3019215
4. Hou, N., Li, Y.: A direct current control scheme with compensation operation and circuit-parameter estimation for full-bridge DC–DC converter. IEEE Transactions on Power Electronics, vol. 36, no. 1, pp. 1130–1142 (2020) doi: 10.1109/TPEL.2020.3002737
5. Villaseñor-Aguilar, M. J., Botello-Álvarez, J. E., Pérez-Pinal, F. J., Cano-Lara, M., León-Galván, M. F., Bravo-Sánchez, M. G., Barranco-Gutierrez, A. I.: Fuzzy classification of the maturity of the tomato using a vision system. Journal of Sensors, vol. 2019, pp. 1–16 (2019) doi: 10.1155/2019/3175848

6. Peña-Aguirre, J. C., Barranco-Gutierrez, A. I., Padilla-Medina, J. A., Espinosa-Calderon, A., Pérez-Pinal, F. J.: Fuzzy logic power management strategy for a residential DC-microgrid. IEEE Access, vol. 8, pp. 116733–116743 (2020) doi: 10.1109/ACCESS.2020.3004611
7. Villaseñor-Aguilar, M. J., Peralta-López, J. E., Lázaro-Mata, D., García-Alcalá, C. E., Padilla-Medina, J. A., Perez-Pinal, F. J., Vázquez-López, J. A., Barranco-Gutiérrez, A., I.: Fuzzy fusion of stereo vision, odometer, and GPS for tracking land vehicles. Mathematics, vol. 10, no. 12, pp. 1–15 (2022) doi: 10.3390/math10122052
8. Zadeh, L. A.: Fuzzy sets. Information and Control, vol. 8, no. 3, pp. 338–353 (1965) doi: 10.1016/S0019-9958(65)90241-X
9. Barranco, A., Cárdenas-León, A., Perez-Pinal, F. J.: Implementación de sistema difuso en arduino uno. Academia Journals Celaya 2016 México (2016)
10. MakeProto: Arduino FIST: MATLAB fuzzy inference system to arduino C converter (2016) http://www.makeproto.com/projects/fuzzy/matlab_arduino_FIST/index.php
11. Khairudin, M., Wijaya, H., Muslikhin, M.: Converter matlab fuzzy inference to arduino C-System. Journal of Physics: Conference Series (ICTVT 2019), vol. 1456 (2020) doi: 10.1088/1742-6596/1456/1/012010

ISSN 1870-4069

Towards an Intelligent Shop Keeper-Centric Transformation

Juan Arturo Pérez-Cebreros[1,2], Gabriel Sánchez-Pérez[1], Luis Mario Hernández-Rojas[1], Ramón Mendoza-Hernández[1], Emiliano Lorences-Gutiérrez[1]

[1] Instituto Politécnico Nacional, ESIME Culhuacán, Departamento de Computación, México

[2] Universidad Jesuita de Guadalajara, Departamento de Electrónica, México

juan.perezc@iteso.mx, gasanchezp@ipn.mx, {lhernandezr1800, rmendozah1600, elorencesg1800}@alumno.ipn.mx

Abstract. Intelligent systems are transforming, in the sense that they are no longer novelties used by some companies, but are now serious business tools that shape the digital world. During the pandemic, different companies sought to digitalize a traditional approach by means of an ecosystem of technological and logistic solutions, whose objective is to strengthen shopkeepers and make them more competitive, applying different models for supply purposes. In this sense, our study is important because it can help grocery wholesalers sustain shopkeepers through face-to-face and digital supply solutions. Adequate categorization of the shopkeeper group is necessary to achieve successful development of marketing strategies that will align perfectly with each group. In this research, we select a Content Management System (CMS) and incorporate a clustering extension with the aim of performing an exploratory analysis of the behavior of different variants pertaining to the K-means algorithm (Hartigan-Wong, Lloyd, Forgy and MacQueen) in the context of a data set of shopkeepers from Mexico City. By means of these Artificial Intelligence techniques, we focus on the prediction of cross-selling opportunities and try to answer questions about who are the most likely shopkeepers to buy additional products, in order to propose an intelligent transformation model focused on the corner shop and study this new way of digitally supplying street stores.

Keywords: Cluster analysis, content management system, shopkeepers.

1 Introduction

During this global pandemic known as COVID-19, shopkeepers or small corner stores, consisting of those small businesses better known as convenience or grocery shops that

are present in residential areas and provide products from the basic food basket, sustain more than 3 million families, are established throughout the country and make large contributions to the national economy.

However, with the arrival of the pandemic, new challenges arose for these small businesses that have put their economic stability at risk.

The importance of digitizing shopkeepers in times of pandemic is demonstrated by emphasizing that these types of stores must take advantage of and update themselves by employing the various digital tools available in order to reach consumers more effectively, as not only can they compete more effectively, but can more easily identify their market, thus precluding that this type of business would close due to the pandemic.

In [1], a platform called "*mandamelo*" is being created for the benefit of corner stores and consumers, making it possible to sell online. In [2] opines that it is necessary for businesses to adapt quickly in order to immediately apprehend the needs of the consumer. For this reason, different ventures have emerged, such as "*Rabbit*" and "*Nutenta*" that offer an application so that you can manage your orders digitally, thus making procedures more efficient, by easing communication and decentralization.

This increase in shopkeeper-centric platforms leads to greater competition, making it necessary to study their different personalities, geographical positions and preferences.

This means that having the same strategy for all shopkeepers is not sufficient [3], so it is vital to be able to categorize them and distinguish the differences between each type of shopkeeper. The aim of this research is to expand Magento in order to focus on finding the most successful groups of shopkeepers among shopkeepers as a whole, while also improving purchase suggestions.

2 Literature Review

The classification and identification of patterns from customer data is very important to support business decision-making [4]. In the marketing context, grouping methods become powerful tools that make it possible to categorize customers in order to identify their patterns and behaviors, in terms of the products or services offered by companies [5].

In [6], the two most popular partition-based clustering algorithms, K-Means and K-Medoids, are evaluated using the transactional dataset. Results from the comparison show that the time spent in the selection of the initial values and the spatial complexity of the cluster overlap is much better in the case of K-Medoids than in the case of K- Means.

Furthermore, K-Medoids is superior in terms of execution time, not sensitive to outliers, and reduces noise compared to KMeans, as it minimizes the sum of the differences of the data objects.

In their research paper, Doğan et al. [7] performed a customer categorization for one of the largest sports retail chains in Turkey, based on the RFM model. The purpose of this research was to find new clusters that would help to redefine the existing card loyalty system, for which they used a data set made up of the Recency, Frequency and Monetary (RFM) indicators for 700032 customers, who made purchases either in person or online.

Two methodologies were applied; firstly, two-stage clustering, from which the Bronze client groups were obtained, with RFM indicators below average; Gold, with an R indicator above average and F and M indicators below, and Premium, with RFM above average; and secondly, clustering by k-means, from which groups of Regular clients emerged, which included 92% of clients and with RFM indicators below average; Loyal, with above-average RFM indicators; Star, with RFM indicators well above average and representing less than 0.015% of customers; and Advanced, with RFM indicators above average but less than those pertaining to the Loyal group.

Tavakoli et al. [8] produced a research article, where the idea was to categorize customers using the K-Means method based on an R + FM model that compared this to traditional RFM, taking into account business changes that make it more effective.

The procedure was applied to Digikala, the largest e-commerce in the Middle East, for which four categories were obtained: assets with high value, medium assets with high monetary value, medium assets with high frequency and low activity value, for which they built and applied marketing strategies for each category.

Results from the various initiatives showed that the new categorization model generated greater impact on customers and therefore had greater effect.

In [9] the importance of cluster analysis in the retail industry is mentioned; in order to identify customers according to their purchasing habits, patterns and behaviors, and thus improve customer service and customer satisfaction, resulting in increased allegiance. K-means clustering results help create a business intelligence system, by providing powerful multidimensional analysis and visualization tools, including building sophisticated data cubes with reference to data analysis needs.

Using a business intelligence application that incorporates this grouping system, as a mechanism to manage a retail business, will provide retailers with the means to categorize customers and better understand their behavior and needs, while making knowledge-based decisions in order to provide a personalized and efficient customer service.

Wu et al. [10] wrote a paper that explains the results obtained, when performing customer categorization in a company dedicated to electronic commerce in Beijing, China. During the investigation, the company's transactional data was analyzed, and then the k-means algorithm fused with the RFM model was applied.

As a result, four clusters categorized according to their purchasing habits were obtained, for which a different CRM strategy was developed in order to achieve a greater level of customer satisfaction. Finally, the company's KPIs improved, verifying the effectiveness of the method applied, as: the number of active customers increased by 519, the purchase volume increased by 279% and total consumption by 102%.

3 Content Management System under Study

The e-Commerce platforms consist of a software system (CMS) that permits you to manage the contents, mainly those of the product to be sold, and through templates provide design to the visual aspect of the online store [11].

Sometimes, depending on the platform chosen, it will not be necessary to have programming knowledge as, thanks to the use of templates, only the specific characteristics required will have to be incorporated.

In this section, we provide details of the selected tools that were studied: Magento, Woocommerce, and Shopify.

a) Shopify is one of the most commonly used ecommerce platforms worldwide, particularly as a result of its versatility when creating a virtual store. Its extensive options for organization and personalization of both products and the store in general, as well as the different processes it offers for payment and the tracking of orders, represent some of its strong points [12].
b) Woocommerce is the WordPress plugin that can turn a website into an eCommerce. Installation is accomplished with a couple of clicks, without requiring knowledge of programming. Once installed, you can add products, create categories or assess shipping costs [13].
c) Magento, in addition to being a recognized CMS or generating web pages, offers its users thousands of alternatives that include methods for payments, shipments, taxes, analytics and logistics; essential for competing in electronic commerce. Basically, it offers a comprehensive open code solution for the development of web pages aimed at online sales; however, for people who lack programming knowledge, Magento is not easy to use [14].

Extending the work in [15], where 36 features were evaluated and compared for each CMS, it was concluded that Magento should receive the best score, followed immediately by Shopify. However, it does not have Artificial Intelligence features that would permit extended cross-selling and up-selling functions. For this reason, in this investigation, we propose an extension to this platform; incorporating automatic learning mechanisms that make an intelligent sale possible.

Importantly, this represents an applied type of research because it intends to solve a real problem and is a non-experimental cross-sectional descriptive design because the data used was limited to a specific period of time and its purpose was to discover customer groups based on their diverse purchasing habits.

The people were all shopkeepers, who made at least one purchase using the e-marketplaces in the study. Likewise, because this project intends to discover hidden patterns based on a considerably large amount of data, 925 customers, who made one or more purchases during the months of May to August 2022 in the e-marketplaces, were selected as a sample.

4 Experimental Procedure

To solve this problem, a three-phase model was proposed:

a) Magento extension phase. In the first phase, we proposed making direct use of the Magento tables using an extraction, transformation and loading process that was carried out in Python.

In Figure 1, we can see part of the Magento data model, which represents a fairly standardized model, but we can access this directly and make use of its master tables and catalog tables, in order to employ data related to customer orders.

This is instead of employing the platform's endpoint or web services because doing so would be impractical and slow.

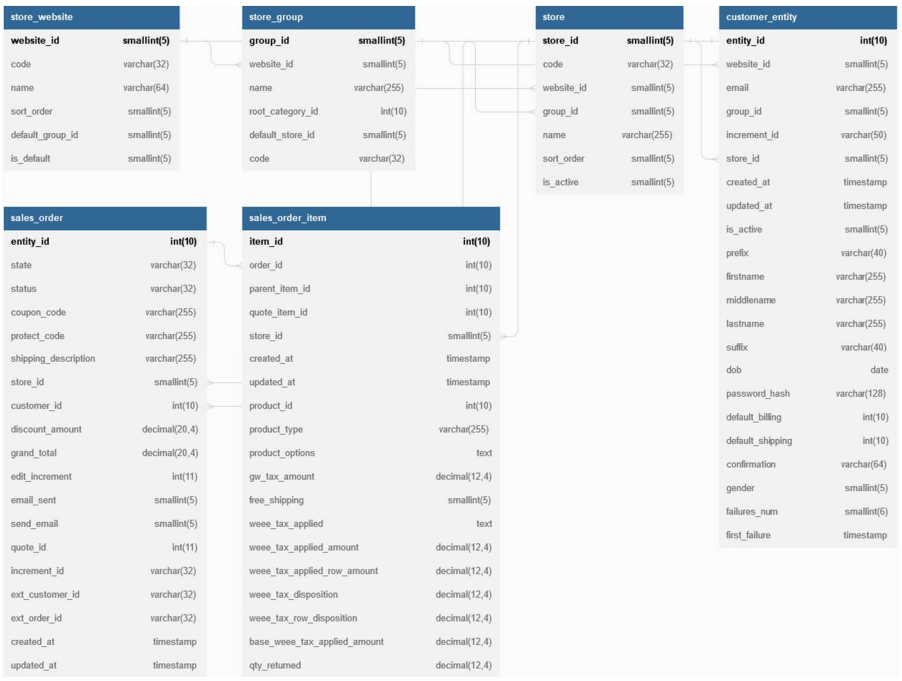


Fig. 1. Diagram for Magento Tables.

The reason for this extension is to provide Magento with intelligent mechanisms for cross-selling and customer categorization, which in our case will focus on shopkeepers. Some previous analyzes [16, 17] indicate that the owners of these convenience stores play a very important role, especially in relation to their personalities.

Therefore, the relationship that is generated between suppliers, customers and their employees promotes commercial development.

b) Shopkeeper categorization phase. Extending the work in [18], we consider the market of a group of shopkeepers based on multiple criteria:

1. $P=\{p_1, p_2,\ldots p_n\}$ is the set of n products.
2. $C=\{c_1, c_2,..c_k\}$ is the set of k evaluation criteria.
3. $T=[t_1,t_2,..,t_j\}$ is the set of j shopkeepers who participate in the market.

These evaluations are used and, with the help of an addition function, a ranking for their products is created. However, this solution is unrealistic due to the large number of shopkeepers and products to be evaluated in the market.

Our proposal is to carry out this ranking automatically. For this, we need to know certain data a priori: we need to know all the products (a*, b*) that compete with each other (this competition will be our criterion).

This will enable us to make comparisons in the form of pairs between each one of the products from the same category (competitive criteria).

To do this, we can make use of the records that are stored in the order tables and automatically establish certain preference rankings for each shopkeeper.

Table 1. Set of comparisons between pairs of products.

Categories	Comparisons between pairs of products
CP_1^{t25}	p11 > p20 > p14 > p34 > p10 > p20 > p32 > p52 > p42 > p63
CP_2^{t25}	p41 > p52 > p56 > p64 > p36 > p77 > p42 > p62
CP_3^{t25}	p67 > p87 > p97 > p66 > p98

To obtain this ranking, we must first define and rank all the comparisons for each product pair (CP) that compete with each other. Let me emphasize that for each $CP(p_i,p_j)_s\ t_r$ where s=1,..,K, in the maximum number of competing subsets (categories). For example, the purchase data for the shopkeeper t_{25} provides us with certain product comparisons (p_i, p_j) and indicates to us that three categories exist, in which we find competing products for sale.

In table 1, we can see the product rankings that result from the comparisons of each pair of products that belong to the same competition criteria. This competition criteria or category may constitute different brands of soft drinks that compete with each other. Thus, the products in a category will be defined as the subset of all the products that compete with each other and for each shopkeeper t_r, we will obtain a subset of possible rankings that represent their preferences.

For example, the set of possible rankings for the shopkeeper t_{25} is made up of three categories and a totality of 23 products that he buys to sell in his store. Therefore, the ranking of each category makes it possible to identify star products and also those least sold for each category.

We propose market categorization based on preferences and direct commercial strategies for each category of shopkeepers. For this purpose, we must first define the degree of similarity in terms of shopkeepers' preferences. This will allow us to establish better trading strategies.

c) Analysis phase for computational experiments. Today, a number of e-commerce companies produce large quantities of data. Most of the time, this data reveals hidden patterns that can be very useful for decision making. One of the ways to acquire new knowledge or find hidden patterns in the data is by applying categorization or clustering algorithms [19]. In [20], shows that the most widely used clustering algorithm is K-means because its results are easy to interpret and there are different implementations.

Something to consider is that in the literature [21, 22, 23], there are data sets where the K-means does not work in the desired way, because the centroids must be adequately separated. For this reason, in our research we carry out a series of experiments with the aim of obtaining a preliminary exploratory analysis of the behavior of different variants among k-means algorithms.

During this phase, we employed R language, which allowed us to implement four variants for this algorithm. These variants are Hartigan-Wong, Lloyd, Forgy, and MacQueen [24, 25]. K-means algorithms include a large number of variants; however, there are few comparative studies.

In this sense, in [25, 26, 27] they implemented variants of these algorithms in Python and R but for synthetic instances in the FCPS repository [28]. In contrast, we select and

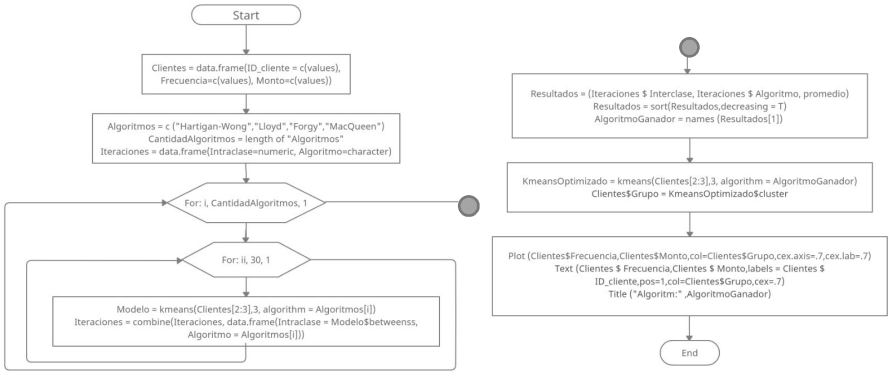


Fig. 2. Diagram for each K-means variant.

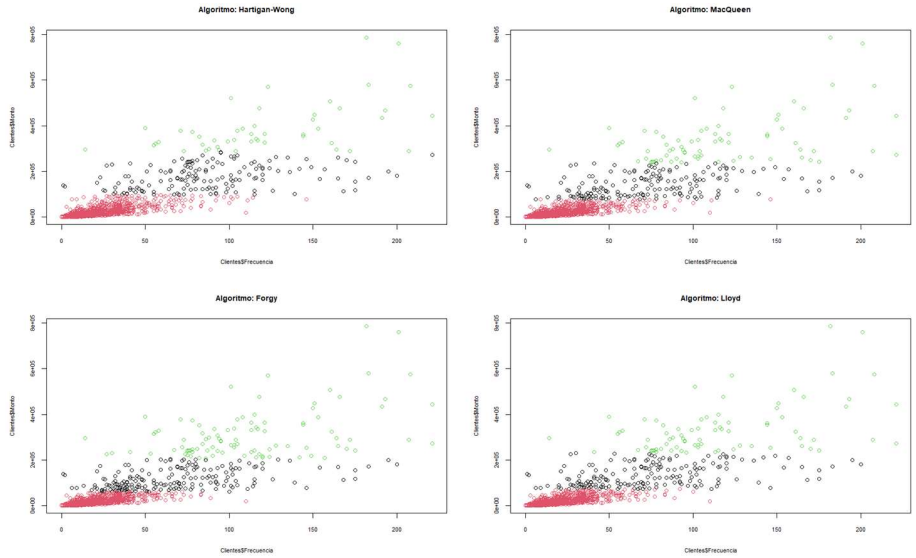


Fig. 3. Graphs showing each of the four variants for the K-means algorithm.

compare the four variants provided by the R language and apply them to our dataset of shopkeepers.

This comparison was made using the intra-cluster distance, which equals the sum of the distances between the centroids. The variant of the algorithm that shows the greatest intra-cluster distance will be the one that best separates the clusters.

Above all, the K-means algorithm is an interactive method that implies dividing a set of n objects into k >= 2 clusters, so that the objects in each cluster are similar to each other but different to the objects in the other clusters [29].

More precisely, the problem that the K-means algorithm solves can be formalized as follows: Let N= $\{t_1,..t_n\}$ where the set of n objects will be the set of shopkeepers, who will be divided and categorized by applying a similarity criterion, where $x_i \in R^d$ for

Table 2. Results from the thirty implementations.

Hartigan-Wong	Lloyd	Forgy	MacQueen
[1] 9.385968e+12	[1] 9.281242e+12	[1] 9.27003e+12	[1] 9.344189e+12
[2] 9.385968e+12	[2] 9.259906e+12	[2] 9.201645e+12	[2] 9.294131e+12
[3] 9.385968e+12	[3] 9.169836e+12	[3] 9.169836e+12	[3] 9.281285e+12
[4] 9.385968e+12	[4] 9.273601e+12	[4] 9.241439e+12	[4] 9.272078e+12
[5] 9.385968e+12	[5] 9.263678e+12	[5] 9.176478e+12	[5] 9.288699e+12
[6] 9.385968e+12	[6] 9.201645e+12	[6] 9.300073e+12	[6] 9.373574e+12
[7] 9.385968e+12	[7] 9.217297e+12	[7] 9.176478e+12	[7] 9.385776e+12
[8] 9.385968e+12	[8] 9.244814e+12	[8] 9.244814e+12	[8] 9.35861e+12
[9] 9.385968e+12	[9] 9.223724e+12	[9] 9.218272e+12	[9] 9.289174e+12
[10] 9.385968e+12	[10] 9.259906e+12	[10] 9.273601e+12	[10] 9.347346e+12
[11] 9.385968e+12	[11] 9.218272e+12	[11] 9.169836e+12	[11] 9.281285e+12
[12] 9.385968e+12	[12] 9.169836e+12	[12] 9.293273e+12	[12] 9.373574e+12
[13] 9.385968e+12	[13] 9.244814e+12	[13] 9.177509e+12	[13] 9.385968e+12
[14] 9.385968e+12	[14] 9.201645e+12	[14] 9.217297e+12	[14] 9.306581e+12
[15] 9.385968e+12	[15] 9.150173e+12	[15] 9.253299e+12	[15] 9.298835e+12
[16] 9.385968e+12	[16] 9.295777e+12	[16] 9.174472e+12	[16] 9.373574e+12
[17] 9.385968e+12	[17] 9.263678e+12	[17] 9.259906e+12	[17] 9.288699e+12
[18] 9.385968e+12	[18] 9.218272e+12	[18] 9.286464e+12	[18] 9.288699e+12
[19] 9.385968e+12	[19] 9.15123e+12	[19] 9.217297e+12	[19] 9.298835e+12
[20] 9.385968e+12	[20] 9.217297e+12	[20] 9.273601e+12	[20] 9.275185e+12
[21] 9.385968e+12	[21] 9.273601e+12	[21] 9.285999e+12	[21] 9.298835e+12
[22] 9.385968e+12	[22] 9.174472e+12	[22] 9.176478e+12	[22] 9.275185e+12
[23] 9.385968e+12	[23] 9.169836e+12	[23] 9.244814e+12	[23] 9.272078e+12
[24] 9.385968e+12	[24] 9.176478e+12	[24] 9.263678e+12	[24] 9.289174e+12
[25] 9.385968e+12	[25] 9.218272e+12	[25] 9.297655e+12	[25] 9.298835e+12
[26] 9.385968e+12	[26] 9.223724e+12	[26] 9.291503e+12	[26] 9.289174e+12
[27] 9.385968e+12	[27] 9.176478e+12	[27] 9.292073e+12	[27] 9.275185e+12
[28] 9.391303e+12	[28] 9.225762e+12	[28] 9.217297e+12	[28] 9.291894e+12
[29] 9.385968e+12	[29] 9.391291e+12	[29] 9.174472e+12	[29] 9.299075e+12
[30] 9.385968e+12	[30] 9.241439e+12	[30] 9.27003e+12	[30] 9.299075e+12

i=1,…, n where d >= 1,…, is the number of dimensions. Additionally, let K>=2 where K must be an integer and K={1,..,k}. For each group k; P={G(1),..,G(k)} of N, let μi be the centroid of the group G(i) for i ∈ K.

The principal reason for forming these groups is to define clusters for the shopkeeper dataset, in such a way that the total intra-cluster distance is minimized. This is achieved by developing the experimental categorization process using variants from the K-means algorithm.

We can observe this in figure 2:

1. We use the data set that will be processed.
2. A variant is selected, and the data is processed.
3. When the algorithm converges, the sum of the intra-cluster distance is stored.
4. Steps a, b and c are repeated 30 times.

5. The average for the intra-cluster distance is calculated, applying the selected variant.

In Table 2, we present the results from each of our thirty implementations, as well as the average intra-cluster distance. Importantly, if the average or all the interactions is considered, the variant that reports the shortest distance is the Forgy.

In all cases the Hartigan-Wong is the one that shows the longest intra-cluster distance. Among the most important results shown for the four variants is the fact that the Hartigan-Wong variant is much better for obtaining the longest intra-cluster distance and on several occasions both Lloyd and Forgy algorithms revealed the shortest intra-cluster distance (15 times).

However, most often it was the Lloyd algorithm that proved optimum for revealing the shortest intra-cluster distance.

5 Conclusions and Future Projects

In this article, we propose an amplification of a CMS such as Magento that will make it possible to incorporate machine learning functions and with this enable an exploratory analysis concerning the behavior of the main variants to the K-means algorithms, applied to a real set of data pertaining to shopkeepers in Mexico City.

In future work, we intend to combine these variants related to clustering algorithms with the proposal we discuss in the shopkeeper categorization phase. With the objective of categorizing the shopkeepers, we provide a ranking of products according to category.

This will inform us in terms of how products compete with each other, and thus show to what extent, the sale of products from different categories of shopkeepers is homogeneous (or heterogeneous). In today's competitive business environment, this will fulfill the great need that grocery wholesalers have to amass, monitor and analyze the data generated by their shopkeepers and competitors.

We are in an era when business represents a war, and like in any war, survival depends on the ability to act quickly in a changing environment. In this sense, we can affirm that the new wave of smart tools will be focused on tracking variables such as operational development, market conditions and the development of competitors, all of them in real time.

However, for many years, the traditional channel has been, as its name indicates, traditional. Despite being the favorite shopping approach throughout Latin America; its processes, mechanisms and tools have been at a standstill for decades. However, due to the pandemic, the digitalization of the corner shop has represented a major challenge for micro-enrepreneurs, in order to adapt to the new requirements of consumers. Smartphones offer a great advantage because they represent the gateway to digitalization.

In our experience, we have seen that some shopkeepers have doubts about how to use tools that are on their smartphones. Often, they do not know how to scan a code or they do not have an email address; these are minor issues but have great impact as it is precisely because of these details that they cannot access digital products and tools that would help them grow their business. If we are proposing the digitalization of small

stores, then the discussion of digital education and training for the use of these tools is essential.

Although there have been real advances towards the digitalization of small stores, great challenges still exist. These businesses are fundamental to daily life; they are supply centers with great tradition and importance to the community. However, for them to continue operating in the long term and to capitalize on real growth opportunities, their gradual digital transformation will be essential.

References

1. Diaz, R., Montalvo, R. F.: Innovative digital transformation strategies of large suppliers for mexican corner stores during a pandemic: challenges and opportunities. Handbook of Research on Digital Innovation and Networking in Post-COVID-19 Organizations, IGI Global, pp. 208–231 (2022) doi: 10.4018/978-1-6684-6762-6.ch011
2. Ziyadin, S., Suieubayeva, S., Utegenova, A.: Digital transformation in business. Lecture Notes in Networks and Systems, Springer International Publishing, pp. 408–415 (2019) doi: 10.1007/978-3-030-27015-5_49
3. Palmatier, R. W., Crecelius, A. T.: The "first principles" of marketing strategy. Academy of Marketing Science Review, vol. 9, no. 1–2, pp. 5–26 (2019) doi: 10.1007/s13162-019-00134- y
4. Kumar, S. J., Oommen-Philip, A.: Achieving market segmentation from B2B insurance client data using RFM and k-means algorithm. In: IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, pp. 463–469 (2022) doi: 10.1109/spices52834.2022.9774051
5. Thomas, J., Preethi, N.: Customer segmentation in the field of marketing. In: 4th International Conference on Recent Trends in Computer Science and Technology, pp. 401–405 (2022) doi: 10.1109/icrtcst54752.2022.9781964
6. Arora, P., Deepali, D., Varshney, S.: Analysis of K-means and K-medoids algorithm for big data. Procedia Computer Science, vol. 78, pp. 507–512 (2016) doi: 10.1016/j.procs.2016.02.095
7. Dogan, O., Ayçin, E., Bulut, Z. A.: Customer segmentation by using RFM model and clustering methods: a case study in retail industry. International Journal of Contemporary Economics and Administrative Sciences, vol. 8, no. 1, pp. 1–19 (2018)
8. Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., Rahmani, R.: Customer segmentation and strategy development based on user behavior analysis, RFM Model and data mining techniques: A case study. In: IEEE 15th International Conference on e-Business Engineering (2018) doi: 10.1109/icebe.2018.00027
9. Anitha, P., Patil, M. M.: RFM model for customer purchase behavior using K-means algorithm. Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 1785–1792 (2022) doi: 10.1016/j.jksuci.2019.12.011
10. Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., Xu, G.: An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means Algorithm. Mathematical Problems in Engineering, vol. 2020, Hindawi Limited, pp. 1–7 (2020) doi: 10.1155/2020/8884227
11. Das, S.: A systematic study of integrated marketing communication and content management system for millennial consumers. Innovations in Digital Branding and Content Marketing. IGI Global, pp. 91–112 (2021) doi: 10.4018/978-1-7998-4420-4.ch005
12. Howe-Patterson, K., Schuiling, I.: Shopify in Germany: An analysis of a Canadian e-commerce platform's marketing strategy and activities in an international market (2020)

13. Sims, L.: Building your online store with WordPress and WooCommerce. Apress (2018) doi: 10.1007/978-1-4842-3846-2
14. Tsagkias, M., King, T. H., Kallumadi, S., Murdock, V., de Rijke, M.: Challenges and research opportunities in eCommerce search and recommendations. Association for Computing Machinery, vol. 54, no. 1, pp. 1–23 (2020) doi: 10.1145/3451964.3451966
15. Abdullah, E. N., Ahmad, S., Ismail, M., Diah, N. M.: Evaluating e-Commerce website content management system in assisting usability issues. In: IEEE Symposium on Industrial Electronics and Applications (2021) doi: 10.1109/isiea51897.2021.9509991
16. Scott, C., Fehlig, V.: Small retailers in/out of the crisis in Germany – Re-inventing retailing. In: SHS Web of Conferences, EDP Sciences, vol. 116., p. 00052 (2021) doi: 10.1051/shsconf/202111600052
17. Hensel, R., Visser, R., Overdiek, A., Sjoer, E.: A small independent retailer's performance: Influenced by innovative strategic decision-making skills? Journal of Innovation and Knowledge, vol. 6, no. 4, pp. 280–289 (2021) doi: 10.1016/j.jik.2021.10.002
18. Liu, J., Liao, X., Huang, W., Liao, X.: Market segmentation: A multiple criteria approach combining preference analysis and segmentation decision. Omega, vol. 83, pp. 1–13 (2019) doi: 10.1016/j.omega.2018.01.008
19. Xian, Z., Keikhosrokiani, P., XinYing, C., Li, Z.: An RFM model using K-means clustering to improve customer segmentation and product recommendation. Advances in Marketing, Customer Relationship Management, and E-Services, IGI Global, pp. 124–145 (2022) doi: 10.4018/978-1-6684-4168-8.ch006
20. Shirazy, A., Hezarkhani, A., Shirazi, A., Khakmardan, S., Rooki, R.: K-means clustering and general regression neural network methods for copper mineralization probability in Char-Farsakh, Iran. Türkiye Jeoloji Bülteni / Geological Bulletin of Turkey, vol. 65, no. 1, pp. 79-92 (2021) doi: 10.25288/tjb.1010636
21. Mittal, M., Sharma, R. K., Singh, V. P.: Performance evaluation of threshold-based and K-means clustering algorithms using iris dataset. Recent Patents on Engineering, vol. 13, no. 2, pp. 131–135 (2019) doi: 10.2174/1872212112666180510153006
22. Fränti, P., Sieranoja, S.: K-means properties on six clustering benchmark datasets. Applied Intelligence, vol. 48, no. 12, pp. 4743–4759 (2018) doi: 10.1007/s10489-018-1238-7
23. Fränti, P., Sieranoja, S.: How much can K-means be improved by using better initialization and repeats? Pattern Recognition, vol. 93, pp. 95–112 (2019) doi: 10.1016/j.patcog.2019.04.014
24. Sreevalsan-Nair, J.: K-Means clustering. Encyclopedia of mathematical geosciences, Springer International Publishing, pp. 1–3 (2021) doi: 10.1007/978-3-030-26050-7_171-1
25. Almanza-Ortega, N. N., Pérez-Ortega, J., Zavala-Díaz, J. C., Solís-Romero, J.: Comparative analysis of K-means variants implemented in R. Computación y Sistemas, vol. 26, no. 1 (2022) doi: 10.13053/cys-26-1-4158
26. Wu, B.: K-means clustering algorithm and python implementation. In: IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (2021) doi: 10.1109/csaiee54046.2021.9543260
27. Thrun, M. C., Ultsch, A.: Using projection-based clustering to find distance- and density-based clusters in high-dimensional data. Journal of Classification, vol. 38, no. 2, pp. 280–312 (2020) doi: 10.1007/s00357-020-09373-2
28. Thrun, M. C., Ultsch, A.: Clustering benchmark datasets exploiting the fundamental clustering problems. Data in Brief, vol. 30, pp. 105501 (2020) doi: 10.1016/j.dib.2020.105501
29. Guo, G., Altrjman, C.: E-Commerce customer segmentation method under improved K-Means Algorithm. In International Conference on Multi-modal Information Analytics, Springer, Cham, vol. 138, pp. 1083–1089 (2022) doi: 10.1007/978-3-031-05484-6_148

ISSN 1870-4069

Suspicious Lung Disease Prediction from Auscultation Sounds Using Neural Networks

Beatriz Anshel Sánchez-García, Said Polanco-Martagón,
Yahir Hernández-Mier, Marco Aurelio Nuño-Maganda,
Jorge Arturo Hernández-Almazán

Universidad Politécnica de Victoria,
Mexico

{1430220,spolancom,yhernandezm,mnunom,
jhernandeza}@upv.edu.mx

Abstract. Early detection of any disease is an important factor in the recovering expectation of any patient. Detection of a disease in an advanced stage can lead to serious or even fatal consequences. To diagnose lung diseases or disorders, physicians use auscultation, which consists of listening body sounds through a stethoscope. This technique requires an outstanding experience of the physician to detect different diseases, and currently artificial intelligence methods are being tested to work as a tool to help in lung disease detection. This article proposes a lung sounds classification method to detect lung issues in suspicious and un- suspicious patients, based on neural networks. This research uses a public dataset, which contains two types of sounds: healthy and lung disease patient sounds. This dataset has a notorious lack of proportion in its data, therefore two balanced techniques were implemented, oversampling and SMOTE, to generate several neural networks models. According to the performed experiments, an accuracy greater than 97% was obtained on the tested dataset.

Keywords: Neural networks, lung sound, classification, oversampling, SMOTE.

1 Introduction

According to [24], respiratory diseases are one of the leading causes of serious diseases worldwide, exceeding 4 million deaths per year. The World Health Organization, in 2017, reported that 10% of the global mortality rate was accounted for by respiratory diseases [21]. Some of these diseases are: Asthma, COPD (Chronic Obstructive Pulmonary Disease), ARI (Upper Respiratory Tract Infection), Bronchiectasis, LRTI (Lower Respiratory Tract Infection), Pneumonia, among others.

Moreover, in 2019 began what would become a global pandemic caused by the virus known as SARS-CoV-2, which is also called COVID-19 [18] and primarily affects the lungs. Hence the importance of detecting any abnormalities in lung sounds.

Physicians commonly use auscultation with a stethoscope as a method to diagnose lung diseases, which is considered to be invasive for the patient. Commonly, the detected sounds are only used during the duration of the examination and are not stored for subsequent analysis.

To be able to use these sounds for research, a medical protocol to generate a dataset would be required, which will result in lots of time and effort, expensive medical equipment and additional expert help in the classification. When acquiring a sound dataset, the stored sound will commonly present noise in the form of digestive sounds, heart sounds, noise due to the stethoscope scraping the skin, among others, hence a cleaning process using electronic or digital filters will be required. These aspects show that the process to generate a dataset of lung sounds is a complicated and expensive process.

In literature, several public datasets can be found [14, 26], which can be down loaded and used freely. However, these datasets have only educational purposes, which implies that in most cases, the data have been extracted and classified for a single purpose and are therefore not proportionally distributed for each category, hindering later data analysis.

Being able to diagnose by auscultation whether or not a patient is suspicious of any pulmonary disease depends heavily on the experience of the physician. In this work, we propose the implementation of a tool to support the diagnosis of lung diagnosis, taking an unbalanced database and processing it to carry out the classification of sounds through a neural network and thus, in a future work, implement it in an embedded system and offer a less invasive solution for the patient.

2 Related Work

In recent decades, lung sounds classification has become a topic of great interest, since it yields relevant information about lung conditions. Several works have been carried out focused on this classification. In [2], a study was conducted between three machine learning approaches to perform classification.

Two of these approaches were based on the extraction of a set of features, which were trained with three classifiers: support vector machines, nearest neighbors and Gaussian mixture models. The third approach is based on convolutional neural networks (CNN). They carried out feature extraction using 12 MFCC (Mel Frequency Cepstral Coefficients) from sounds and the local binary pattern extracted from spectrograms.

They used the dataset of the R.A.L.E. [29] (Respiration Acoustics Laboratory Environment) which contained approximately 50 recordings of respiratory sounds. This dataset required a prior request to the author to grant permission to use the data, it is not a public dataset. In this same work [3], the authors carried out a data augmentation using spectrogram clipping techniques and vocal tract length perturbation. Subsequently, they classified sounds into 7 different classes: normal, coarse crackle, fine crackle, monophonic wheeze, polyphonic wheeze, squawks and stridor.

They obtained 91.12% accuracy using support vector machines with the MFCC and 95.96% accuracy with CNN, after performing 1 million iterations. Although CNNs are one of the most widely used models for classification, it is important to note that its performance depends largely on the learning parameters, on whether the dataset is large and on the number of iterations carried out to train it, which can be time consuming and requires significant computational resources. On the other hand, in [17], authors performed a multichannel classification using a recurrent convolutional neural network. Using their proposed device, they obtained lung sounds from 16 healthy people and from 7 people diagnosed with idiopathic pulmonary fibrosis (IPF), obtaining a total of 23 sounds in total.

Subsequently, they extracted features from the spectrogram to later classify as healthy or pathological (binary classification), using an MLP (Multilayer Perceptron), a BiGRNN (Bidirectional Gated Recurrent Neural Network) and a ConvBiGRNN (Convolutional Bidirectional Gated Recurrent Neural Network).

The latter being the one that obtained the best results with F1 = 92.4%. However, the small number of patients with IPF, the large age difference between healthy patients and patients with IPF, the little variety in terms of lung conditions, among others, would cause the models not to have a correct training process.

On the other side, in [7], the authors carried out a multiclass classification: normal, asthma, heart failure, pneumonia, bronchiectasis and bronchitis and chronic obstructive pulmonary disease. They obtained 308 sound recordings and 1176 from the ICBHI (International Conference on Biomedical and Health Informatics Challange) dataset [14], the same dataset that was used in this study, performed this data fusion to solve the unbalance problem that arose.

They obtained 98% accuracy by making use of the entropy features with Powered Decision Trees. Although his technique works and his results are high, they do not exceed those obtained in this study. In other research [13], a telemedicine framework was built to predict respiratory pathology by lung sound examination. They compared all three approaches with machine learning for lung sound detection. The proposed telemedicine framework trained through Bagging and Boosting classifiers, Improved Random Forest, AdaBoost and Gradient Boosting algorithm. Using a set of handdrawn features they obtained 95% of accuracy.

In another article [30], authors used a dataset composed of five types of lung sounds: normal, coarse crackle, fine crackle, monophonic and polyphonic wheezing. They used higher order statistics (HOS) to extract features, second, third and fourth order accumulators, and genetic algorithms (GA), achieving 96.9% of accuracy. However, these last two required a long training time (between 30 and 200 minutes), which can be considered as a disadvantage.

Based on the results found in the literature, some of the methodologies and tools used in the related works can be useful. Table 1 summarizes the information related to this work that was found in the literature.

The main aspects that were used to make this Table are: the used method, the used datasets, the classification method and their results. It can be seen that the previously proposed models have already reached optimal values for classification. However, the results can be considered unreliable, since they may show an optimal value but perform part of the classification in an erroneous way, since they present an imbalance in their data.

3 Objective

The objective of this research is to train a multilayer perceptron to predict whether a patient is suspected or not suspected of lung disease through the classification of lung sounds using a public database by applying data balancing techniques.

Table 1. Summary background.

Ref	Dataset	Balanced dataset	Clasification	Multi-channel/ single-channel	Method	Sounds	Metric
[17]	Own	Yes	Binary	Multi-channel	ConvBiGRNN	Healthy and idiopathic pulmonary fibrosis	F1 = 92.4%
[17]	Own	Yes	Binary	Multi-channel	BiGRNN	Healthy and idiopathic pulmonary fibrosis	F1 = 86.1%
[17]	Own	Yes	Binary	Multi-channel	MLP	Healthy and idiopathic pulmonary fibrosis	F1 = 50.2%
[7]	Own and public (The same)	Most of the Classes	Multiclass	Single-channel	Boosted Decision Trees	Healthy, asthma heart failure pneumonia, bronchitis and chronic obstructive pulmonary disease	Accuracy = 98.27%
[7]	Own and public (The same)	Most of the Classes	Multiclass	Single-channel	Linear Discriminat	Healthy, asthma heart failure pneumonia, bronchitis and chronic obstructive pulmonary disease	Accuracy = 86.41%
[7]	Own and public (The same)	Most of the Classes	Multiclass	Single-channel	Support Vector Machine	Healthy, asthma heart failure pneumonia, bronchitis and chronic obstructive pulmonary disease	Accuracy = 98.20%
[7]	Own and public (The same)	Most of the Classes	Multiclass	Single-channel	K-Nearest Neighboors	Healthy, asthma heart failure pneumonia, bronchitis and chronic obstructive pulmonary disease	Accuracy = 97.04%
[13]	Own and public	Not mentioned	Binary	Single-channel	SVM	Coarse ralescence, crackles and wheezing	Precision = 81.0%
[13]	Own and public	Not mentioned	Binary	Single-channel	KNN	Coarse ralescence, crackles and wheezing	Precision = 94.1%
[13]	Own and public	Not mentioned	Binary	Single-channel	Naive Bayes	Coarse ralescence, crackles and wheezing	Precision = 81.0%
[30]	Public	Not mentioned	Multiclass	Single-channel	Improved-RF	Healthy, coarseness, coarse crackles, monophonic wheezing monophonic wheezing stridor and squawking	Accuracy = 98.76%
[30]	Public	Not mentioned	Multiclass	Single-channel	AdaBoost	Healthy, coarseness, coarse crackles, monophonic wheezing monophonic wheezing stridor and squawking	Accuracy = 96.29%
[30]	Public	Not mentioned	Multiclass	Single-channel	Gradient Boosting	Healthy, coarseness, coarse crackles, monophonic wheezing monophonic wheezing stridor and squawking	Accuracy = 94.71%
[2]	Public	No	Multiclass	Single-channel	Ensembling CNN	Healthy, coarse crackles, fine crackles, monophonic wheezing, poliphonic wheezing stridor and squawking	Accuracy = 95.56%
[2]	Public	No	Multiclass	Single-channel	MFC-SVM	Healthy, coarse crackles, fine crackles, monophonic wheezing, poliphonic wheezing, stridor and squawking	Accuracy = 91.12%
[2]	Public	No	Multiclass	Single-channel	LBP-SVM	Healthy, coarse crackles, fine crackles, monophonic wheezing, poliphonic wheezing, stridor and squawking	Accuracy = 71.21
[2]	Public	No	Multiclass	Single-channel	MFCC-CNN	Healthy, coarse crackles, fine crackles, monophonic wheezing, poliphonic wheezing, stridor and squawking	Accuracy = 91.67%
[2]	Public	No	Multiclass	Single-channel	LBP-SVM	Healthy, coarse crackles, fine crackles, monophonic wheezing, poliphonic wheezing, stridor and squawking	Accuracy = 80.00%

Table 2. Computer specs.

	Specification
Operating system	Windows 11
System	64 bits
Processor	Intel(R) Core(TM) Graphics 3.70 GHz
RAM	8 GB
Hard drive	SSD 494.5 GB

4 Methodology

4.1 Dataset

In order to carry out the training of a neural network, it is necessary to have a large enough dataset to ensure optimal learning. In this work, a public dataset [14] was used, which was developed by two research teams in Portugal and Greece. It contains 920 recordings (900 of sick people and 20 of healthy people) which were obtained from 126 people and its duration varies between 20 and 90 seconds.

These sounds were obtained with the following devices: AKG C417L Micro- phone (AKGC417L), 3M Littmann Classic II SE Stethoscope (LittC2SE), 3M Litmmann 3200 Electronic Stethoscope (Litt3200), WelchAllyn Meditron Mas- ter Elite Electronic Stethoscope (Meditron), and the age of the patients who took part of the tests where children, adults and elderly.

The dataset is organized as follows:

- 920 .wav sound files.
- 920 .txt files.
- A file that organizes the diagnostic of each patient.
- A file that describes the elements are contained in the name of each sound.
- A file that organizes the 91 sound names from dataset.
- A file that describes demographic information from each patient.

The sounds were taken from 7 different positions on the chest: trachea, an- terior left, anterior right, rear left, rear right, lateral left and right lateral, and they include clean breath sounds as well as noisy sounds. The diseases found in the dataset are Asthma, COPD (Chronic Obstructive Pulmonary Disease), ARI (Upper Respiratory Tract Infection), Bronchiectasis, LRTI (Lower Respiratory Tract Infection) and Pneumonia, which in this work were grouped in the “suspicious” class.

Used Hardware: This section gives details about the hardware that was used during the development and implementation of the proposed model. Table II gives details of the computer equipment that was used.

4.2 Feature Extraction

In the tests performed, five features were extracted from each sound, the most commonly used in the literature were chosen. Each sound consists of feature vectors of length 24, including the MFCC (20), i.e. the four chosen features plus the 20 MFCCs. These features are described below:

Zero Crossing Rate: As shown in Figure 1, zero crossing rate is presented as the average of the number of times the signal switches between positive and negative within the time window. The speed at which these crossings occur is considered as a measure of the frequency content of a signal. It is commonly used in speech recognition and music information restoration [23, 32].

Spectral Centroid: It is a characteristic that details the timbre of a sound approaching the perceptive brightness it possesses, in addition, as shown in Figure 2, it is the one in charge of pointing out where the "center of mass of the sound" is located and it is presented as the result of the weighted average of its frequencies. In the same way, it can be considered as the center of gravity spectrum frequency components have. It is worth mentioning that this does not remain fixed but varies and consequently, the features also change [20, 15].

Spectral Decrease: It is defined as the measure of the asymmetry that the spectral shape of a signal possesses, as shown in Figure 3, it represents the frequency that is below a punctual percentage of the total spectral energy. Also, it is known as the *N* percentile of the magnitude distribution of the spectrum and its values are usually 85 or 95 percent [4, 31, 8].

Mel Frequency Cepstral Coefficients (MFCCs): These coefficients are a set of between 10 and 20 features that help to show the shape of a spectral envelope. Obtaining these coefficients is one of the most efficient and important techniques of voice parameterization, it also uses the Fourier Transform to carry out the obtaining of the frequencies of a signal. Its main purpose is to obtain a consistent, complete and adequate representation in order to achieve a statistical model of the sound with a greater degree of precision [11]. In Figure 4 we can see an example obtained from a lung sound.

Chroma Features: They are defined as a type of musical scales, as shown in Figure 5, the spectrum is drawn in 12 containers of sounds or notes, these have intervals that are always at the same distance and are equivalent to 12 semitones (also called chroma) belonging to the musical octave [27, 22].

4.3 Neural Networks

Artificial neural networks are machine learning techniques that simulate the learning of biological organisms. The strengths of synaptic connections commonly occur randomly in response to external stimulation, this is how learning occurs in living organisms. This biological mechanism is simulated in neural networks, which contain calculation units called neurons.

The computational units are connected to each other through weights, which fulfill the same function as the strengths of synaptic connections in biological organisms. In a few words, they are like a type of machine learning algorithm, which is based on the behavior of neurons in the brain. In this work, a MLP will be used to carry out the classification of lung sounds.

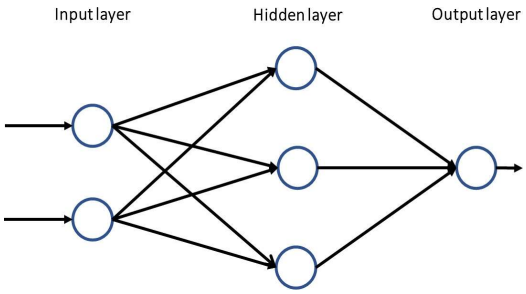


Fig. 1. Diagram of the structure of a two input layer MLP.

Multi-layer Perceptron (MLP): The MLP [9] is one of the simplest, most common and widely used deep learning models of neural networks. They are also called feedforward since the flow of information through the network is forward, that means, there is no feedback.

The perceptron consists of a simple mathematical function where a set of inputs is entered, some mathematical operations are carried out and a result of said operation is obtained. The equation representing the perceptron is shown in equation 1:

$$y = \Sigma(w_i * x_i + c), \tag{1}$$

where, w_i represents the input weights, x_i the perceptron inputs and c the activation function.

MLP composition: MLP architecture is made up from the following elements [28]:

- An input layer, commonly identify as x.
- A number of hidden layers.
- An output layer, commonly identify as y.
- A group of weights and biases for each layer, commonly identify as w y b.
- An activation function for each hidden layer.

Figure 2 shows the architecture of a two-layer MLP (it is important to mention that it is very common for the input layer to be excluded from the total number of layers of the neural network):

Equation 2 shows the computation of the output layer, y, of an MLP. This output layer is illustrated in Figure 6:

$$y = \sigma(W_2\sigma(W_1x + b_1) + b_2), \tag{2}$$

where, y represents the output, W the weights and b, the biases. We can observe in equation 2, that the only variables that affect the result obtained in the output 'y', are the weights (W) and the biases (b). That is why it is so important that these variables have the correct values, in order to determine the best possible prediction. This process of adjusting the values is known as training a neural network [25].

If we look at the process of each of the interactions that occur in the network layers, it can be found that it is divided into 2 events [16]:

Table 3. Activation functions.

Function	Equation
Identity	$f(x) = x$
Sigmoid	$f(x) = \frac{1}{}$
Tanh	$f(x) = tanh(x) = 2 - 1$
ReLU	$f(x) = \{Oforx \leq 0$
	$f(x) = \{Oforx \geq 0$

The computation of outputs from the input layer to the output layer, also called Feedforward.

An update of the values of the weights and biases, this process is known as backpropagation.

Activation Functions: When designing a neural network model, it is required to define which activation function will be used. This function has an important role, since they are the ones in charge of taking an input value and transforming it to move it to the next layer [10]. The Table 3 shows the activation functions used in this work.

Data Balancing Techniques: In this study an unbalanced dataset was used, thus two techniques were used to better distribute the data among the classes.

Oversampling: This technique is responsible for balancing the distribution of data by duplicating the elements of the minority class. However, examples containing noise are generated which could lead to problems when training a classification model [3].

SMOTE: It is a technique that performs an oversampling of the minority class in order to increase the number of elements and get a more balanced dataset. Unlike OverSampling, SMOTE does not make copies of minority elements, but rather takes features of these elements and their neighbors to subsequently generate new elements with a combination [3].

4.4 Results Comparison

For our case of classification, we were used only two possible variables: suspicious or unsuspicious, therefore, we were used a confusion matrix with 2 class labels. Each of the predictions can be one of four outcomes shown in the confusion matrix:

- True Positive (TP): True prediction and actually true.
- True Negative (TN): False prediction and actually false.
- False Positive (FP): True prediction and actually false.
- False Negative (FN): False prediction and actually true.

In addition, four metrics were implemented that were the basis for obtaining the results, which are described below:

Precision: It is responsible of measuring the quality of the machine learning model in classification tasks, the lower the dispersion of the data, the greater the precision. It is the

Table 4. Neural Network Topologies used in this work.

Test number	Layers	Neurons
Test 1	2	128, 64 respectively
Test 2	2	50, 60 respectively
Test 3	4	60, 70, 80, 90 respectively
Test 4	4	50, 60, 70, 80 respectively
Test 5	4	50, 60, 90, 100 respectively

result of dividing the true positives by all the positive results (including both the true and fake ones). Briefly, it is the percentage of positive cases obtained by the model [28].

The precision is given by equation 3:

$$Precision = \frac{TP.}{TP+FP}. \quad (3)$$

Recall: Also known as the True Positive Rate, it is which informs us of the number of positive cases that were correctly identified by the model [28]. It is calculate as shown in equation 4:

$$Recall = \frac{TP.}{TP+FN}. \quad (4)$$

F1-Score: It is the result of the combination of precision and recall in a single value, which makes it practical, since it is much easier to compare the performance of the model [28].

F1 is computed by taking the harmonic mean between precision and recall, as shown in equation 5:

$$f1 = 2 \times \frac{precision \times recall}{precision+recall}. \quad (5)$$

With these metrics, four possible cases can be obtained for each class [1]:

High precision and low recall: The model has difficulties detecting the class, but when it does, it is reliable.

High precision and high recall: The model has an excellent handling of this class.

Low precision and high recall: The model detects the class well, but tends to include samples that do not belong to that class.

Low precision and low recall: The model cannot correctly classify the class.

When working with a dataset that presents imbalance, a high precision value is regularly obtained in the majority class and a low value in the minority class. These cases frequently occur in the health area, which is why in this work we used data contrasting techniques.

Accuracy: Shows the percentage of cases that the model managed to get right [28].

It is computed through equation 6:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (6)$$

It is the most used measure to evaluate the quality of the models, taking values between 0 and 1. The closer it is to 1, the better.

Table 5. Evaluation metrics results (Unbalanced data).

Test number	Accuracy	Recall	F1 score	Precision
Test 1	0.95%	0.98%	0.97%	0.97%
Test 2	0.97%	1.00%	0.98%	0.97%
Test 3	0.96%	0.97%	0.98%	0.98%
Test 4	0.95%	0.98%	0.97%	0.96%
Test 5	0.97%	0.98%	0.98%	0.98%

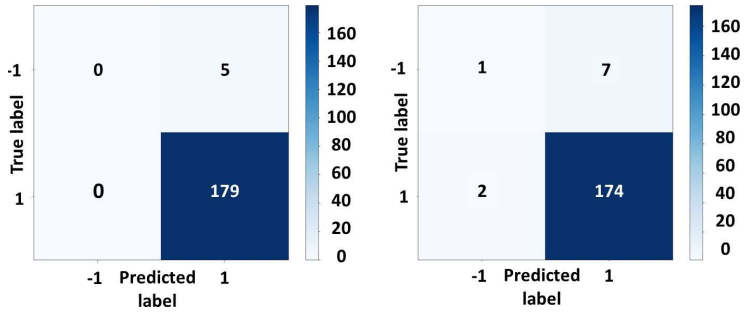


Fig. 2. Confusion matrix set one (test 2 vs test 4).

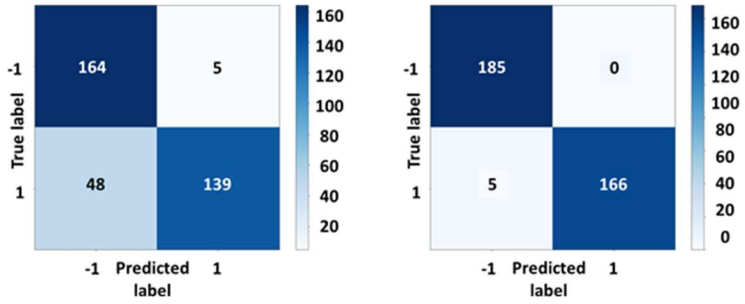


Fig. 3. Confusion matrix set two (test 2 vs test 3).

Table 6. Evaluation metrics results (OverSampling).

Test number	Accuracy	Recall	F1 score	Precision
Test 1	0.96%	0.93%	0.96%	1.00%
Test 2	0.85%	0.74%	0.83%	0.96%
Test 3	0.98%	0.97%	0.98%	1.00%
Test 4	0.97%	0.95%	0.97%	1.00%
Test 5	0.94%	0.90%	0.94%	1.00%

5 Results

After training the neural network, 3 different sets of tests were developed, with a total of 5 tests each. The model was validated through the use of Cross Validation. As can be seen in Table 4, five different network topologies were used for each test, these were chosen

Table 7. Evaluation metrics results (SMOTE).

Test number	Accuracy	Recall	F1 score	Precision
Test 1	0.94%	0.91%	0.93%	0.96%
Test 2	0.89%	0.78%	0.87%	0.99%
Test 3	0.97%	0.95%	0.97%	1.00%
Test 4	0.98%	0.97%	0.98%	1.00%
Test 5	0.94%	0.88%	0.93%	1.00%

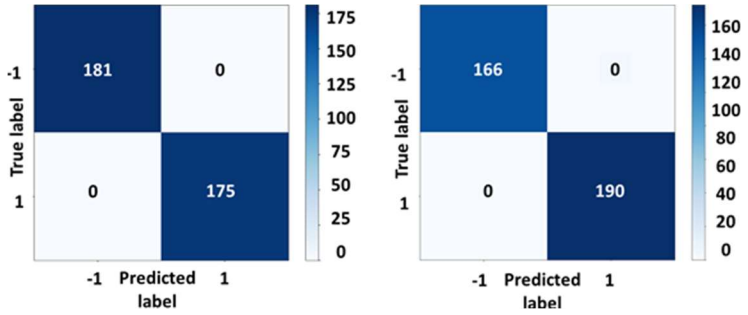


Fig. 4. Confusion matrix set one (test 4 vs test 2).

after several tests where it was obtained that the best results were obtained with networks that had 2 to 4 hidden layers and 50 to 100 neurons each.

In the first set of tests, the training and classification proceeded with unbalanced data, as can be seen in Table 5, the results obtained contain percentages that indicates that the classification is being carried out correctly.

However, Figure 2 shows us that despite the metrics indicating that an optimal classification was being generating, this was only being done for the positive class, which means that the model was not learning the negative class, since the dataset was unbalanced, That was the main reason why it was decided to use techniques that would help countervail the data.

To improve the results from the first set of tests, OverSampling technique were used for training and classifying the second set, which was used to balance datasets class.

As in test set 1, the evaluation metrics showed high percentages (a supposed indication of a correct classification). However, it can be seen in Figure 3 (test 3 vs. test 2) that the OverSampling technique was not having the expected results. The reason was that the technique used only duplicated the minority class sounds, therefore the neural network was not optimally trained. And although in test 3, the metrics and the confusion matrix showed good results, the training could not be fully trusted, since it was being carried out with duplicate data. As a solution, in the third set of tests the technique was replaced by SMOTE, which, unlike Oversampling, did not duplicate the sounds, but took one of them and generated a new one applying a minimal change.

The results of applying this technique are shown in Table 7. It can be seen in Figure 4 (test 4 vs test 2) that the classification result into values of 0, which corresponds to the cells of values classified in an erroneous way. This shows that the training of the network had achieved its objective and the model had a greater control over the classes. Based on the evaluation metrics and the confusion matrix, it was found that the test that obtained

the best results was test 4 of set 3 (SMOTE technique) with 98%, 97%, 98% and 100% accuracy, exhaustiveness, F-value and precision respectively.

6 Conclusion

In this work, a method for lung sound classification based on Neural Networks is presented. It is important to note that if a dataset is not balanced, it will be difficult to generate models to correctly predict an outcome. That is why the contribution of this work is considered to be the use of data balancing techniques, which showed a considerable improvement in the results obtained by the model.

The classes used for this work were divided into 10 versus 90. When the proposed techniques were used, 200% of artificial data were obtained, yielding 50% vs. 50%. The classes that were used for this work were divided into 10% vs 90%. When the proposed techniques were used, 200% of artificial data were obtained. In the performed experiments, the SMOTE technique achieved the best results, getting a 98.97% of accuracy.

As a future work, it is intended to develop a digital stethoscope for the automatic detection of a probable lung disease, that is why this simple classification model is used, since the implementation would be much simpler due to its low processing cost and resources. It should be noted that the proposed model is just a tool to support the detection of pulmonary diseases and does not intend to replace the diagnosis of a health expert.

References

1. Aggarwal, C. C.: Neural networks and deep learning: A textbook. Springer (2018) 10.1007/978-3-319-94463-0
2. Bardou. D., Zhang, K., Ahmad, S.: Lung sounds classification using convolutional neural networks. Artificial Intelligence in Medicine, vol. 88, pp. 58–69 (2018) doi: 10.1016/j.artmed.2018.04.008
3. Zhang, J., Bloedorn, E., Rosen, L., Venese, D.: Learning rules from highly unbalanced data sets. In: Fourth IEEE International Conference on Data Mining, pp. 571–574 (2004) doi: 10.1109/ICDM.2004.10015
4. Cord M., Cunningham, P.: Machine learning techniques for multimedia: Case Studies on Organization and Retrieval. Springer (2008) doi: 10.1007/978-3-540-75171-7
5. Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., Fan, H.: Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks, vol. 130, pp. 22–32 (2020) doi: 10.1016/j.neunet.2020.06.015
6. Er, M. B.: Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features. Applied Acoustics, vol. 180, pp. 108152 (2021) doi: 10.1016/j.apacoust.2021.108152
7. Fraiwan, L., Hassanin, O., Fraiwan, M., Khassawneh, B., Ibnian, A. M., Alkhodari, M.: Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers. Biocybernetics and Biomedical Engineering, vol. 41, no. 1, pp. 1–14 (2021) doi: 10.1016/j.bbe.2020.11.003
8. Theodoros, G., Pikrakis, A.: Introduction to audio analysis (2014) doi: 10.1016/ C2012-0-03524-7
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press (2016) www. deeplearningbook.org

10. Gavin, H.: Mastering machine learning with scikit-learn: Apply effective learning algorithms to real-world problems using scikit-learn (2014)
11. Abdul, Z. K., Al-Talabani, A. K.: Mel frequency cepstral coefficient and its applications: A Review. IEEE Access, vol. 10, pp. 122136–122158 (2022) doi: 10.1109/access.2022.3223444
12. Nor-Syazwani, I.: Speech recognition using Mel frequency cepstral coefficient (MFCC). Universiti Tun Hussein Onn Malaysia, 2010
13. Jaber, M. M., Abd, S. K., Shakeel, P. M., Burhanuddin, M. A., Mohammed, M. A., Yussof, S.: A telemedicine tool framework for lung sounds classification using ensemble classifier algorithms. Measurement, vol. 162, pp. 107883 (2020) doi: 10.1016/j.measurement.2020.107883
14. Kaggle.: Respiratory Sound Database. English (2019) www.kaggle.com/datasets/vbookshelf/respiratory-sound-database
15. Lokki, T., Müller, M., Serafin, S., Välimäki, V.: Sound and music computing. Multidisciplinary Digital Publishing Institute (2018) doi: 10.3390/books978-3-03842-908-1
16. Loy, J.: Neural network projects with python. Packt Publishing Ltd (2019)
17. Messner, E., Fediuk, M., Swatek, P., Scheidl, S., Smolle-Jüttner, F. M., Olschewski, H., Pernkopf, F.: Multi-channel lung sound classification with convolutional recurrent neural networks. Computers in Biology and Medicine, vol. 122, p. 103831 (2020) doi: 10.1016/j.compbiomed.2020.103831
18. Gobierno de México. COVID-19. Español (2021) coronavirus.gob.mx/covid-19/
19. Millstein, F.: Deep learning: 2 manuscripts-deep learning with Keras and convolutional networks in python (2018)
20. Park, T. H.: Introduction to digital signal processing. World Scientific (2009) doi: 10.1142/6705
21. Perna, D.: Convolutional neural networks learning from respiratory data. In: IEEE International Conference on Bioinformatics and Biomedicine (2018) doi: 10.1109/bibm.2018.8621273
22. Poynton, C.: Digital video and HD: algorithms and interfaces (2007)
23. Lawrence, R. R.: Introducing to digital speech processing. Publishers Inc (2007)
24. Naves, R., Barbosa, B. H. G., Ferreira, D. D.: Classification of lung sounds using higher-order statistics: A divide-and-conquer approach. Computer Methods and Programs in Biomedicine, vol. 129, pp. 12–20 (2016) doi: 10.1016/j.cmpb.20 16.02.013
25. Raschka, S., Liu, Y. H., Mirjalili, V.: Machine learning with PyTorch and Scikit-Learn (2022)
26. E-learning resources: Reference database of respiratory sounds-wheezes (2022) www.ers-education.org/e-learning/reference-database-of-respiratory-sounds/wheezes
27. Taha-Sencar, H., Memon, N.: Digital image forensics: There is More to a Picture than Meets the Eye (2012) doi: 10.1007/978-1-4614-0757-7
28. Siahaan, V., Hasiholan-Sianipar, R.: Step by step tutorials on deep learning using scikit-learn, Keras, and tensorflow with python GUI (2021)
29. The R.A.L.E. repository. English (2008) www.rale.ca/lungsounds.html
30. Ulukaya, S., Serbes, G., Kahya, Y. P.: Resonance based separation and energy based classification of lung sounds using tunable wavelet transform. Computers in Biology and Medicine, vol. 131, pp. 104288 (2021) doi: 10.1016/j.compbiomed. 2021.104288
31. Zelkowitz, M. V.: Advances in computers vol. 44 (2010)
32. Zhang, T., Kuo, C. C. J.: Content-based audio classification and retrieval for audiovisual data parsing. Springer US (2001) doi: 10.1007/978-1-4757-3339-6

ISSN 1870-4069

Prediction of the Polarity of Opinions in the Domain of Tourism through Machine Learning

Marcos A. Leiva-Vasconcellos, Mireya Tovar-Vidal

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

marcos.leiva@alumno.buap.mx,
mireya.tovar@correo.buap.mx

Abstract. The analysis of the polarity of any type of comments has increased thanks to the development of Web 2.0 where millions of opinions are currently generated by users of various sites, with high information content. Opinion mining focuses on automatically determining the polarity of publications for research and development of real-world applications. This article aims to determine which of the proposed algorithms (Decision Trees, Support Vector Machine and *Naïve Bayes*) are appropriate for predicting the polarity of opinions in the tourism domain, for this a set of opinions (487) about hotels are extracted from TripAdvisor. The experimental results obtained show that Support Vector Machine (SVM) and *Naïve Bayes* are the best classifiers for this type of task with an accuracy of 62% and 61% respectively, a result that will improve by increasing the training set.

Keywords: Opinion mining, natural language processing, machine learning.

1 Introduction

With the creation of Web 2.0, the user went from being a consumer of resources to a content creator, having the possibility of issuing criteria and evaluating the content on the Internet. TripAdvisor was created in 2000 and although it was not intended for users to exchange opinions about the sites visited.

From 2004, consumer comments exceeded professional comments, it became a collection of comments from travelers from around the world where the individual experience of the places visited was exposed, according to a radio interview with Stephen Kaufer, creator from TripAdvisor.

Fig. 1 shows the number of views in millions from 2014 to 2020 from the TripAdvisor.com website, confirming that the website is the largest travel guide in the world. According to [9, 4], 1 in 16 people consult TripAdvisor.com to plan their vacation, which is why the opinions posted on the site are so important to the tourism sector.

According to the World Tourism Organization, tourism is defined as that which includes the activities conducted by people during their trips and stays in places other

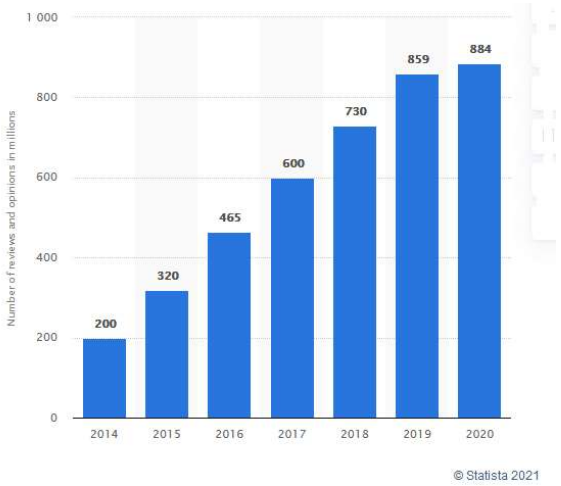


Fig. 1. Number of reviews, expressed in millions, on TripAdvisor. Source: Statista.com.

than their usual environment, for a consecutive period of less than one year for leisure purposes, for business and others [6].

Tourism is one of the fundamental activities in many countries, including Mexico, which represented 8.7% of its GDP (Gross Domestic Product) at the end of December 2019, according to [8], which is why it is constantly receiving feedback to improve the quality of tourism services. It is of vital importance, due to the importance that tourists give it, to consult the opinions that are published by the people of the tourist places to improve their quality.

Sentiment analysis or opinion mining is a field of research within Natural Language Processing that automatically extracts subjective information expressed in a text about a given domain [7]. In this way, the author's attitude about a particular topic can be known, which can normally be positive, neutral, or negative. The interest in opinion mining has increased over time due to the large amount of information that circulates in the networks and that it would be impossible for a person to characterize it [22]. In this sense, the tourism sector can rely on opinion mining to automatically extract the polarity of tourists.

Studies have been conducted using sentiment analysis to assess opinions about a topic. Since TripAdvisor is the largest site for tourism opinions, it has been used as a data source to develop algorithms that evaluate the opinions of users regarding sites, and thus be able to influence problems in a practical way.

Evaluating sentiment in large volumes of data is not always unambiguous, even when done manually the result may not be the same, depending on the encoder. This article aims to determine the appropriate algorithms for prediction of opinion polarity in the tourism domain using opinion mining. The study is carried out using TripAdvisor as a data source.

This article is structured as follows: Section 2 present the works related. Section 3 shows the proposed solution of the identification of entities. Section 4 shows the results obtained, and in Section 5 the conclusions.

2 Related Works

Opinion mining has been investigated for some years using Natural Language Processing, below are some works related to this study.

In [3] association rules are applied to the database, in this case Twitter, with these rules, opinions can be categorized, and people's feelings identified. In the research work developed by [16], the authors propose a model composed of three phases: Pre-processing, Identification of Aspects and Polarity Identification, obtaining 50% effectiveness in the SemEval 2016 Competition forum. They use the sentiment dictionary SentiWordNet, which is generally so that some words in specific contexts have one polarity and, in another context, a different one, is used for comments in Spanish.

In the research [1] they implement a scheme for the unsupervised detection of the polarity of opinions from new lexical resources SentiWordNet 4.0 and 4.1 obtaining values of accuracy and *F1* of 85% much higher than version 3.0.

The author in [12] proposes a model based on 3 modules: text processing, attribute selection and classification with machine learning, extracting comment data from TripAdvisor, Booking, Expedia and Trivago. These comments are classified into 2 classes (good and bad), the classifiers used were Support Vector Machine (SVM), *Naïve Bayes* and decision trees, the experiments showed that *Naïve Bayes* was the most accurate, although the accuracy levels are accurate for SVM and *Naïve Bayes*.

In [19] the author uses lemmatization and normalization to train the proposed model, which uses SVM, which makes the classification set obtain better results and is done for comments in Spanish.

The authors in [2] extract the keywords from the comment to obtain lists of concepts and keywords through the Microsoft Knowledge Graph. In the research [17] the authors use 3 classifiers to evaluate comments on Twitter, they are: Decision trees, *k nearest* neighbors and *Naïve Bayes*, the best classifier of all was the decision tree.

Sentitext is used in [11], which is a sentiment analysis system, based on domain-independent linguistic knowledge using the Freeling morphological analyzer. It uses comments in Spanish from TripAdvisor and has a success rate close to 90%, although it detects more positive segments than negative ones. This work performs a sentiment analysis classification using the main classifiers according to previous studies to determine which ones provide the best results.

3 Proposed Solution

Next, the proposed solution is described, as well as the methodology to follow for the prediction of the polarity of opinions in the tourism field.

3.1 Description of the Proposed Solution

After reviewing the main trends in sentiment analysis, it can be said that more research is aimed at classifying opinions as positive, neutral or negative.

Most are in the English language and very few in Spanish. The most used classifiers are SVM and *Naïve Bayes*, giving good results in each of them.

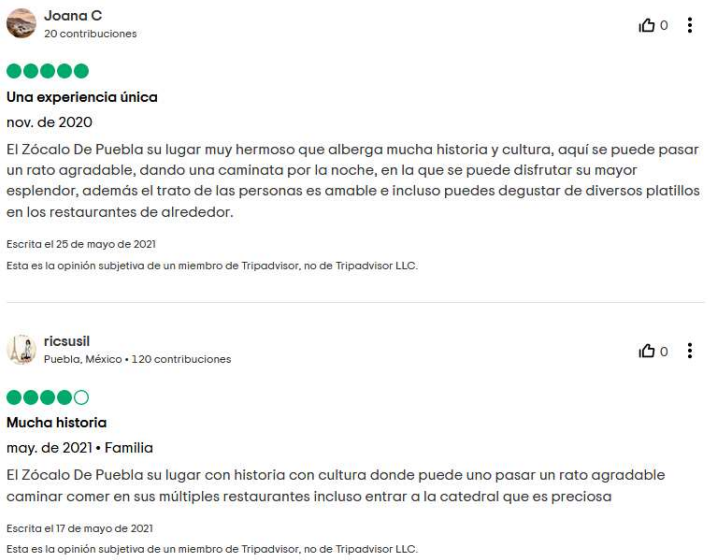


Fig. 2. Reviews on TripAdvisor.com. Source: TripAdvisor.com

There are not many studies of natural language processing using TripAdvisor, most of the research focuses on Twitter and Facebook. The consulted investigations do not consider the date that the opinion was made to verify if the comment was improved over time.

Our proposal consists of 5 types of polarities like the rate on TripAdvisor and are described below:

1. Very Negative: It refers to negative opinions, but with emphasis, it uses expressions such as: very bad, worse, etc.
2. Negative: Use negative expressions, for example: bad, expensive, etc.
3. Regular: They are regular expressions such as: we went to the hotel, room on the 4th floor, etc.
4. Positive: Expressions with an emphasis on positive issues: good hotel, nice pool, etc.
5. Very Positive: They are positive expressions with emphasis: very good, very nice, etc.

To carry out the program, it was determined to use Python due to its robustness in terms of natural language processing through the NLTK (*Natural Language Toolkit*) library, which, although it was not initially designed for the Spanish language, can use the corpus in Spanish as es cess_ esp[10], which has 500,000 words and 610 files[15]. The data structure to be used will be trees, graphs and json mainly for the representation of information.

Reviews are sourced from TripAdvisor.com in Spanish language; Fig. 2 shows the form in which the comments are expressed, according to [20] and [21], users sometimes

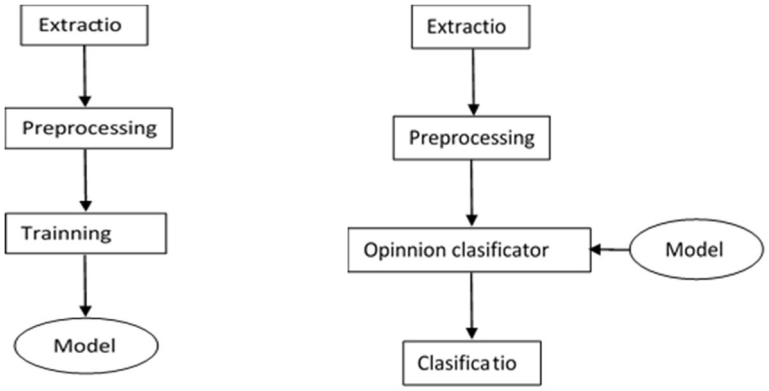


Fig. 3. Methodology for the classification of opinions a) Training b) Classification with the generated model.

evaluate in one way and the opinions expressed contradict the evaluation given, so it is not feasible to take only the numerical evaluation of each user. *Web scraping* is used to obtain the comments of a certain place since TripAdvisor does not have an API to obtain the data.

3.2 Proposed Methodology

The methodology to be used for polarity prediction is described below. Fig. 3 a) shows the methodology for training the classifiers, returning the model to be used in the classification and figure 3 b) shows the methodology for performing the classification according to the model obtained in the training. Next, the stages in each of the steps will be described.

3.2.1 Extraction

The data extraction will be done from the TripAdvisor.com page using web scraping, through a script implemented in Python and the data will be saved in csv files.

3.2.2 Preprocessing

In this stage, a set of techniques will be applied to obtain better results in the later stages, the Python NLTK (Natural Language Toolkit) library is used. Characters that are not letters will be removed first, such as: punctuation marks, numbers and characters that do not belong to the Spanish alphabet. In addition, pieces of the text that may interfere with the analysis of the text will be removed; this is domain dependent. In this phase, all words are converted to lowercase. Finally, the noise will be eliminated, which consists of getting rid of stopwords (words like *the, the, are*, etc.).

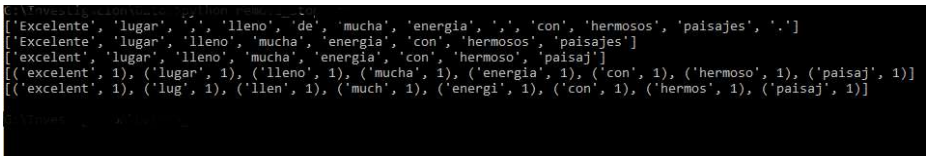


Fig. 4. Pre-processing using the Python NLTK library.

For example, if you have the opinion: "Excellent place, full of energy, with beautiful landscapes. "When applying these steps, the text would read as follows: "Excellent place full of energy with beautiful landscapes". Then it would go to the tokenization process, which consists of separating the words from the text and building a vector composed of each word.

The last method of pre-processing is called lemmatization or stemming, which reduces the original word in its root part, making subsequent classification easier. Continuing with the previous example, after applying tokenization and lemmatization, the result would be as follows: [('excellent', 1), ('lug', 1), ('llen', 1), ('much', 1), ('energi', 1), ('con', 1), ('hermos', 1), ('landscape', 1)], in Fig. 4 shows the whole process using the aforementioned library with *Snowball Stemmer*.

3.2.3 Training

For this phase, *TF-IDF* are used, which are: Term Frequency and Inverse Frequency of Documents and with this it will convert the document feature vectors using TfidfVectorizer. The vectorized document is used for SVM [18], *Naïve Bayes* [13] and Decision Tree [14] training to generate a model, which will be loaded into the classification.

3.2.4 Opinion Classifier

There are various investigations about classifiers for opinion mining, in most investigations machine learning techniques are used, being SVM and *Naïve Bayes* the most popular, of these two techniques SVM is the one that presents greater certainty according to [5]. Due to the aforementioned, it is determined to use SVM and *Naïve Bayes* as opinion classifiers, although it will be trained and classified using Decision Trees to verify its results.

3.3 Evaluation of the Proposal

To measure the performance of a classifier, several terms have been defined that are described below [1]:

Accuracy : *is* the proportion of the total number of predictions that were correct as shown in equation 1:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}. \quad (1)$$

Precision: is the proportion of predicted cases that were positive as shown in equation 2:

$$P = \frac{TP}{TP+FP}. \quad (2)$$

Recall: is the proportion of positive cases that were correctly identified as shown in equation 3:

$$R = \frac{TP}{TP+FN}. \quad (3)$$

F_1: is the harmonic mean that combines the precision and accuracy values as shown in equation 4:

$$F_1 = \frac{2*P*R}{P+R}. \quad (4)$$

Being:

- *TP* (*True Positive*) True Positive: Number of cases that the test declares positive and that are truly positive.
- *TN* (*True Negative)* True Negative: Number of cases that the test declares negative and that are negative.
- *FP* (False *Positive*) False Positive: Number of cases that the test declares positive and that are negative.
- *FN* (False *Negative*) False Negative: Number of cases that the test declares negative and that are positive.

4 Experimental Results

This section describes the data used for the analysis of the opinions and the results corresponding to each class with the SVM, *Naïve Bayes* and Decision Trees algorithms.

4.1 Dataset

The data used is extracted from the TripAdvisor.com page by web scraping 45 hotels in Puebla. Table 1 shows the names of the hotels.

Table 2 shows the total opinions classified by classes from 1 (very negative) to 5 (very positive), out of a total of 487 opinions retrieved from the aforementioned hotels.

4.2 Data Set

With the SVM and *Naïve Bayes algorithms* presented in section 3 and adding decision trees, the training and classification are carried out. Table 3 shows the results of Accuracy, Precision, Recall and F_1 for each of the algorithms. The training was carried out with cross validation, in which 75% of the cases were taken for training and the remaining 25% of tests. The Baseline was created with a Random Forest Classifier using a random class to the testing set.

Table 1. Selected Hotels for Feedback.

Casona Maria Boutique Hotel	New Suite Seville	Hotel Posada Cuetzalan
La Quinta by Wyndham Puebla Palmas Angelopolis	Hotel Royal 500	Hotel Panamerican
Hotel Casona Poblana	Moctezuma Luxury Boutique Hotel B&B	San Jose House of Prayer Hotel
Palm House Hotel	Hotel San Miguel	Palace Hotel
Loa_Inn_Puebla	Meson San Sabastian	Hotel Royalty Center
Hotel_Gilfer	Hotel Leones	eight 30
Hotel_One_Puebla_FINSA	Hotel Real Santander	Isabel Hotel
Blue Talavera Hotel	Hotel Las Iglesias	Sonata Hotel & Residences
Fiesta Inn Puebla Finsa	OYO Hotel Del Paseo	Puebla de Antano
Hotel Puebla Plaza	Puebla American Party	Crowne Plaza Puebla
Hotel Star Express	Diana Hotel	Suites La Concordia
El Capricho Boutique Hotel	poblana square	Portal Hotel
Hotel Hacienda Del Molino	Meson Sacristy of the Company	Palacio Julio Hotel
Hotel La Quinta	Hotel Casona San Antonio	Hotel Gilfer
Hotel Plaza Zacapoaxtla	Quinta Esencia Hotel Boutique	Loa Inn Juarez

Table 2. Distribution of opinions by classes (1 to 5).

	1	**2**	**3**	**4**	**5**
Opinions	30	13	42	129	273

Table 3. Results of the experiment using the 3 algorithms mentioned.

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1_**
Naïve Bayes	61.48	1.0	0.61	0.76
SVM	62.30	0.88	0.62	0.71
Decision tree	48.36	0.47	0.48	0.47
Baseline	22.13	0.89	0.22	0.34

5 Conclusions and Future Work

This article presents automatic classification algorithms that allow identifying the polarity of the opinions extracted from TripAdvisor. The polarity is distributed from 1 (very negative) to 5 (very positive).

Even though the training set is very small, the experimental results show that the SVM method achieves good results, obtaining 62%, as well as *Naïve Bayes,* which has 61% accuracy, all the algorithms are above the baseline created. We can see that the decision tree algorithm does not have a good performance (48%) as reflected in the studies consulted.

It has been shown that the SVM and *Naïve Bayes algorithms* are appropriate for polarity prediction in this domain, so it is proposed as future work to enrich the data set. In addition to implementing other artificial intelligence techniques such as Artificial Neural Network to improve the performance.

References

1. Amores-Fernández, M., Arco, L., Borroto, C.: Unsupervised opinion polarity detection based on new lexical resources. Computación y Sistemas, vol. 20, no. 2, pp. 263–177 (2016) doi: 10.13053/cys-20-2-2318
2. Chen, W., Xu, Z., Zheng, X., Yu, Q., Luo, Y.: Research on sentiment classification of online travel review text. Applied Sciences, vol. 10, no. 15 (2020) doi: 10.3390/ app10155275
3. Díaz-García, J. Á., Ruiz, M. D., Martín-Bautista, M. J.: Minería de opinión no supervisada en Twitter. In: XVIII Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2018) Granada, Spain, pp. 1023-1028 (2018)
4. Filieri, R., Acikgoz, F., Ndou, V., Dwivedi, Y.: Is TripAdvisor still relevant? The influence of review credibility, review usefulness, and ease of use on consumers' continuance intention. International Journal of Contemporary Hospitality Management, vol. 33, no. 1, (2020) doi: 10.1108/IJCHM-05-2020-0402
5. Flores, L., Guadalupe, I., Peña-Álvarez, E. P.: Aprendizaje automático para la optimización de procesos de marketing digital en el sector turístico. Universidad Tecnológica de Perú (2020)
6. Ghanem, J.: Conceptualizing "the Tourist": A critical review of UNWTO definition. Master Thesis, Universitat de Girona (2017)
7. Hariguna, T., Sukmana, H. T., Kim, J.: Survey opinion using sentiment analysis. Journal of Applied Data Sciences, vol. 1, no. 1, pp. 35–40 (2020) doi: 10.47738/jads.v1i1.10
8. Instituto nacional de estadística y geografía (INEGI): Turismo. Sistema de Cuentas Nacionales de México (2021) https://www.inegi.org.mx/temas/turismosat/#Informacion _general
9. Kinstler, L.: How TripAdvisor changed travel. The Guardian, London (2018) https://www.theguardian.com/news/2018/aug/17/how-tripadvisor-changed-travel
10. Martí, M., Taulé, M., Recasens, M.: AnCora: Multilevel annotated corpora for Catalan and Spanish. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08) (2008)
11. Moreno-Ortiz, A., Pineda-Castillo, F., Hidalgo-García, R.: Análisis de valoraciones de usuario de hoteles con Sentitext: Un sistema de análisis de sentimiento independiente del dominio. Procesamiento del Lenguaje Natural, no. 45, pp. 31–39 (2010)
12. Mostafa, L.: Machine learning-based sentiment analysis for analyzing the traveler's reviews on Egyptian hotels. In: Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), pp. 405–413 (2020) doi: 10.1007/978-3-030-44289-7_38
13. Murphy, K. P.: Naive Bayes classifiers. University of British Columbia (2006)
14. Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., Brown, S. D: An introduction to decision tree modeling. Journal of Chemometrics, vol. 18, pp. 275–285 (2004) doi: 10.10 02/cem.873
15. Navas-Loro, M., Rodríguez-Doncel, V.: Spanish corpora for sentiment analysis: A survey. Language Resources and Evaluation, vol. 54, pp. 303–340 (2020) doi: 10.1007/s10579-019-09470-8
16. Rosales-Quiroga, M. A., Vilariño-Ayala, D., Pinto, D., Tovar, M., Beltrán, B.: Análisis de sentimientos basado en aspectos: un modelo para identificar la polaridad de críticas de usuarios. Research in Computing Science, vol. 115, pp. 171-180 (2016)
17. Shoeb, M., Ahmed, J.: Sentiment analysis and classification of tweets using data mining. International Research Journal of Engineering and Technology (IRJET), vol. 4, no. 12 (2017)
18. Suthaharan, S.: Support vector machine. Machine learning models and algorithms for big data classification, Springer, pp. 207-235 (2016) doi: 10.1007/978-1-4899-7641-3

19. Ticona-Nina, R.: Minería de opiniones basado en aprendizaje supervisado en la evaluación de destinos turísticos de la región de Puno. Universidad Peruana Unión (2019)
20. Valdivia, A., Luzón, M. V., Herrera, F.: Sentiment analysis in TripAdvisor. IEEE Intelligent Systems, vol. 32, no. 4, pp. 72–77 (2017) doi: 10.1109/MIS.2017.3121555
21. Valdivia, A., Luzón, M. V, Herrera, F.: Sentiment analysis on TripAdvisor: Are there inconsistencies in user reviews? In: International Conference on Hybrid Artificial Intelligence Systems, vol. 10334, pp. 15–25 (2017) doi: 10.1007/978-3-319-59650-1_2
22. Vazquez, K. L., Tovar, M., Vilariño, D., Beltrán, B.: Un algoritmo para detectar la polaridad de opiniones en los dominios de laptops y restaurantes. Research in Computing Science, vol. 128, pp. 91–98 (2016) doi: 10.13053/rcs-128-1-8

ISSN 1870-4069

Upgrading Relations List in Fuzzy Cognitive Maps Using Reinforcement Learning

Frank Balmaseda, Mabel Frias, Frank Verstappen

Hasselt Universiteit Martelarenlaan,
Belgium

{frank.balmaseda, mabel.friasdominguez,
frank.verstappen}@uhasselt.be

Abstract. Fuzzy Cognitive Maps (FCM) are dedicated to modeling complex dynamic systems and has been widely studied. One of those studies probed that Computing with Words (CWW) is very effective to improve the interpretability and transparency of FCM. Learning methods to calculate the weight matrix in a map are the target of hundreds of studies in various parts of the world. These methods allow the map to learn or evolve towards a better state, but always taking into account that the output must be evaluated and compared using a real scenario. Reinforcement Learning has, within its performance, one of the possible answers to this problem, aimed at improving the classification capacity of the map and thus improving the learning of its relations list. This paper presents a new learning method for Fuzzy Cognitive Maps. The proposal was evaluated using international databases and the experimental results show a satisfactory performance.

Keywords: Fuzzy cognitive maps, reinforcement learning, relations list.

1 Introduction

Cognitive maps were presented for the first time in 1976 by [3], their main objective being to represent social scientific knowledge through designated digraphs where the arcs are causal connections between the nodes. Fuzzy Cognitive Maps (FCM) [19], were introduced as an extension of Cognitive Maps theory, and they are recurrent structures modeled through weighted graphs.

In this representation, each node or concept of the graph may represent a variable, an object, entity or state of the system that is intended to be modeled; while the weights in the connections determine the causality between these concepts.

There are two fundamental strategies to build FCMs: manual and automatic. In the first variant, the experts determine the concepts that describe the system, as well as the direction and intensity of the causal connections between the neurons.

In the second variant, the topology of the map and the weight matrix are constructed from historical data, or representative examples that define the behavior of the system. Regardless of the strategy, domain experts need to define the architecture that best fits the system that is intended to model, as well as the restrictions inherent in the model [21].

The formulation of learning algorithms to estimate the causality of the system (matrix of causal weights that regulate the interaction between the concepts) is still an open problem for the scientific community [23]. The learning algorithms that are dedicated to the classification of the weights in FCM still present gaps, both in the efficiency and in the results that they offer.

For instance, in [5] authors recognize that using Ant Colony System outperforms RCGA, NHL and DD-NHL algorithms in terms of model error, but only when multiple response sequences are used in the learning process, not so when one response sequence is used.

Other studies like in [17] and [26], shows the proposals are a little bit slow because for computing every weight it is necessary to consider the other concepts involved in the causal-effect relation for a target concept. The broad attention of the authors to the subject reinforces the premise of its importance.

Within this framework, there is a clear need to improve the methods used (or, as is the case of this paper, to create new ones) up to now to carry out the learning of the FCM.

2 Fuzzy Cognitive Maps and Computing with Words

In a FCM the inference can be defined mathematically using a state vector and a causality matrix. The state vector A_{1xN} represents the activation degree of the concepts, and the causal weight matrix W_{NxN} defines the interaction between the concepts.

The activation of C_i will depend on the activation of the neurons that directly affect the concept C_i and the causal relationships associated with that concept. The process of inference in an FCM is then summarized in finding the value of the state vector A through time for an initial condition A^0 as can be seen in the equation (1):

$$A^{(t+1)} = f\left(\sum_{j=1.}^{N} A^{(t)} W_{ji}\right), i \quad j \tag{1}$$

The causal relationships between the concepts can occur in 3 ways. For two concepts C_i and C_j it is fulfilled that:

- $W_{ij} > 0$: Indicates a positive causality between the concepts C_i and C_j, which means an increase (decrease) in the value of C_i leads to the increase (decrease) in the value of C_j.
- $W_{ij} < 0$: Indicates a negative causality between the concepts C_i and C_j, that is, the increase (decrease) in the value of C_i leads the decrease (increase) in the value of C_j proportional to the absolute value of $W_{ij.}$
- $W_{ij} = 0$: Indicates the non-existence of a causal relationship between C_i and C_j. It can also occurs when W_{ij} is very close to zero.

On the other hand, CWW [30] was presented as a methodology that allows introducing linguistic variables with the objective of performing computational operations with words instead numbers. Linguistic variables describe situations that cannot be clearly defined in quantitative terms and allow a very particular type of

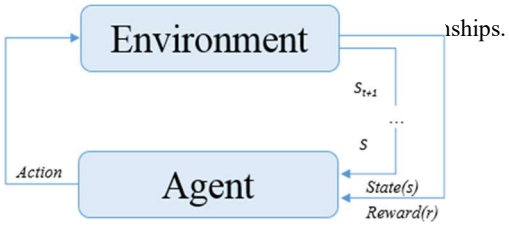


Fig. 1. Behavior of a reinforcement learning problem.

translation from natural language to numerical or numerical sentences. Some models inspired in his technique are briefly described below:

- *Linguistic Computational Model based on membership functions.* The linguistic terms are expressed by fuzzy numbers, which are usually described by membership functions. This computational model makes the computations directly on the membership functions of the linguistic terms by using the Extension Principle [9].
- Linguistic Computational Symbolic Model [6]. This model performs the computation of indexes attached to linguistic terms. Usually, it imposes a linear order to the set of linguistic terms $S = S_0, ..., S_g$ where $S_i < S_j$ if and only if $i < j$.
- *The 2-tuple Fuzzy Linguistic Representation Model* [15]. The above models perform simple operations with high transparency, but they have a common drawback: the loss of information caused by the need of expressing results in a discrete domain. The 2-tuple model is based on the notion of symbolic translation that allows expressing a domain of linguistic expressions as a continuous universe.

3 FCM and Reinforcement Learning

The main problem of reinforcement learning methods is to find an optimal policy of actions that is capable of reaching a goal maximizing the rewards. Starting from an initial state, the agent chooses an action in each iteration, which is evaluated in the environment and receives penalties or rewards according to the results [27].

The Figure 1 shows a general performance of this technique.

The learning methods for FCM in the literature formulate several hypotheses and most of them have application results that improve the classification of FCM. Below, some of these approaches are briefly described:

- *Differential Hebbian Learning* [8] is based on the following principle: if the cause concept C_i and the effect concept C_j change the activation value simultaneously, then the causal weight W_{ij} will increase with a constant factor, otherwise the causality will not be modified in that iteration, see equation 2:

$$w_{ij}^{(t+1)} = \begin{cases} W_{ij}^{(t)} + \gamma^{(t)}[\Delta A_i/\Delta A_j - W_{ij}^{(t)}] & if\ \Delta A_i \neq 0 \\ W_{ij}^{(t)} & if\ \Delta A_j = 0 \end{cases} \quad (2)$$

- (*Balanced Differential Algorithm* [17] proposes to use the activation degree of the nodes that are simultaneously modified during the weight matrix updating. This algorithm assume it is a FCM without concepts using auto-connections and that it is adjusted iteratively.

- In [20] the authors propose a method based on Evolutionary Strategy to adjust the structure of the map using historical data, where each instance is composed of a pair of vectors. The first vector encodes the activation degree of the input concepts, while the second vector denotes the response of the system for the specified configuration.

- In [24] the authors applied Particle Swarm Optimization to estimate the appropriate structure of the system from a sequence of multiple state vectors. The peculiarity of this method lies in its learning scheme: the response of the system is defined by decision-making nodes.

- A new automated approach based on Genetic Algorithms with Real Coding can be found in [26]. The idea of the method is to maximize the objective function $(x) = 1/(1 + \alpha\ (x)$, reducing the global error of the map.
- In [5] the authors presented a novel algorithm inspired in Ant Colony Optimization to adjust maps with dozens of concepts. In this model, each weight is encoded as a sequence of discrete states. During the search process, each ant constructs a solution. In the next step, each discrete solution is trans- formed to its continuous equivalent, resulting in a matrix of causal weights.

3.1 Learning Algorithm

The method proposed in [10] uses a Learning Automata (LA) [22, 28], which are simple reinforcement learning components for adaptive decision making in unknown environments. An LA operates in a feedback loop with its environment and receives feedback (reward or punishment) for the actions taken. The general update scheme is given by 3 and 4:

$$p_m(t+1) = p_m(t) + \alpha_{reward}(1 - \beta(t))\ (1 - p_m(t)) - \alpha_{penalty}\beta(t)\ p_m(t), \quad (3)$$

if a_m is the action taken at time t:

$$p_j = p_j(t+1) - \alpha_{reward}(1 - \beta(t))\ p_j(t) + \alpha_{penalty}\beta(t)\ (r-1)^{-1} - p_j(t), \quad (4)$$
$$if\ a_m \neq a_j$$

Algorithm 1 Establishing classification

1: Create map (concepts and relationships). Executed by an expert
2: Assign matrix weights to the map using FCM+CWW+RL. Executed by expert
3:Assign to each concept C_i, the value of each feature of the object to be classified.
4: Assign the linguistic term "NA" to each decision concept.
5: Calculate the activation value of the concepts
6: Classify a new object O_c using equation 5.

Algorithm 2 FCM+CWW+RL

1 : $\forall\, i = 1...n, \forall\, j = 1...n, M_{ij} = 1/n$
T2: Obtain reward $r_i =$
3: Update M using r_i.
4: Repeat from step two until reach the stopping condition

where $p_i(t)$ is the probability of selecting action i at time step t. The constant α_{reward} and $\alpha_{penalty}$ are the reward and penalty parameters. When $\alpha_{reward} = \alpha_{penalty}$, the algorithm is referred to as linear reward-penalty (L_{R-P}), when $\alpha_{penalty} = 0$, it is referred to as linear reward-inaction (L_{R-I}) and when $\alpha_{penalty}$ is small compared to α_{reward}, it is called linear *reward ε penalty* ($L_{R-\varepsilon P}$). $\beta(t)$ is the reward received by the reinforcement signal for an action taken at time step t. r is the number of actions [29].

In [10] authors propose the use of RL in decision making problems specifically in Personnel Selection. Basing in this approach, in this paper it is proposed using RL as a learning method to adjust the weight matrix of a FCM+CWW. To reach this objective define a FCM+CWW as the environment over which the agent must make the decisions. The parameters to be used in the environment are defined as follows:

- Firstly, the size of the list of terms that are used in the map: For this proposal a study is included in [11] that sates, through a transfer function, the use of linguistic terms instead numerical methods traditionally used to rep- resent the weights in the causal relationships of the maps. This list may vary depending on the needs of precision in the domain, therefore, it can be simplified to three terms (Low, Medium, High), however, the author suggests to use fifteen terms as described in [11]
- Length of the relations list: A FCM contains a list of relationships used to determine the connections between each concept involved in the map.
- Training set: These instances are necessary for the classification algorithm to train with that set and to return the best individual. The training set is created from the cross validation technique used to perform the data validation.

From this point, the environment continues with obtaining the reward and updating the weight matrix, as is shown in algorithm 1:

$$O_c = F(\arg \max_{A_k \in A^D} \{A_k\}), \tag{5}$$

Table 1. Matrix built from the list of relationships.

R1	R2	R3	R4	R5
R_{i_1}	R_{i_2}	R_{i_3}	R_{i_4}	R_{i_N}
R_{j_1}	R_{j_2}	R_{j_3}	R_{j_4}	R_{j_N}
W1	W2	W3	W4	W5

where F returns the value of the class D_k and A^D is the activation set of the decision concepts. It is worth mentioning that our symbolic model based on FCM conserves its recurrent nature.

This implies that the FCM will produce a state vector composed by linguistic terms in each iteration until it discover a fixed point or reach a maximum number of iterations. Next step will be building a Nx4 matrix using the relations list. In Table 1 is shown a generic matrix of the relations list which describes the behavior of each relation. Here R1 is a relation, R_{i_1} is the starting concept of R1 and R_{j_1} is the target concept of R1. In this approach, the main idea is not to add relations to the list, but using the initial list of relations, change the R_i and R_j and keeping the same weight, which will be estimated using [4].

This method suggests using the classification accuracy value as a reward. This value tends to 1 when the reward is high. In equation (6) we formalize the expression used to calculate the reward once the classification is done, where r_i is the reward in the iteration, δ_i is the number of well-classified objects for the same iteration, and T is the total instances of the data set:

$$r_i = \frac{\delta_i}{T}. \tag{6}$$

With this value, the agent performs the updating of the weight matrix and the reward is calculated in each iteration, which substitutes the time T commonly used for this type of problem. Below, the steps of the method are detailed presented in algorithm 2.

Modifying the reward function on each iteration allows finding a new state of the map and evaluating it in order to determine improvements in the classification, otherwise, this state will not be visited again through the iterations.

4 Experimental Results

This section presents the experimental framework that uses 20 data sets from [2]. The cross-validation method [7] was used to validate the results that subdivided the data set into *k* subsets of equal size (in this experimental frame k = 10), of which one part makes up the set of tests and the other part the set of training. When using this technique, the number of subsets with which we worked was taken into account, because with higher values for *k*, the trend of the real error range of the estimate is reduced.

Table 2. Results of accuracy for each method.

Data-sets	MLP	NB	SVM	RF	ID3	FCM+RL
apendicitis	82,09	85,09	86,49	80,26	70	87,41
blood tranfussion	75,81	75,81	56,25	62	75,14	76,03
chocardiograme	90,02	90,82	90,69	65,83	85,43	91,21
heart-5an-nn	72,95	81,8	82	74,2	65,55	83,03
iris	94,67	94,67	82,22	70,35	89,33	96,66
liver disorders	53,59	62,88	65,38	66	49,82	66,13
planing-relax	63,27	60,94	73,22	58,75	53,27	73,37
saheart	67,11	67,97	87,18	63,99	49,57	67,99
yeast3	91,3	91,29	93,34	92,36	88,8	93,42
pima 5an-nn	67,03	74,22	75,78	74,91	59,51	76,09
acute inflammation	100	100	100	99,87	100	100
Yield	51,78	62,5	65,34	63,8	53,57	66,48
phoneme	81.1	76,62	82	79,77	80,6	83,07
ecolio	81,87	68,51	81,6	64,88	62,5	82,01
iris-5an-nn	96,42	95,33	96	64,88	62,5	82,01
vehicle	68,72	54,48	70,68	59,32	58,75	70,93
wine	96,81	94,97	97	92,12	91,54	96,8
glasso	95,56	94,28	95,75	89,41	96,3	96,46
vertebra-column-3c	82,54	66,24	83,95	74	65,83	84,86
car	72,83	66,27	74,99	71,38	63,99	77,83

Table 3. Holm test for α=0.05 for classification accuracy, taking as control FCM+CWW+RL.

i	Algorithms	$z = \frac{(R0 - Ri)}{SE}$	p	Holm Null	Hypothesis
5	ID3	4.60014	0.000033	0.0004	Rejected
4	RF	4.35842	0.005008	0.01	Rejected
3	SVM	3.11039	0.005201	0.01	Rejected
2	MLP	2.62014	0.008776	0.025	Rejected
1	NP	1.71063	0.010283	0.05	Rejected

In addition, the estimate will be more precise, the variance of the error range real is greater and the necessary computation time also increases as the number of experiments to be performed increases. For the statistical analysis were used the hypothesis test techniques in [25, 13], Friedman and Iman-Davenport for multiple comparisons [18] in order to detect statistically significant differences in a group of the results.

Holm test [16] is used in these experimental studies to find significantly higher algorithms. These tests are suggested in [7, 14, 13, 12], where the authors agree in the relevance of using them for validating results on the reinforcement learning area. KEEL module Non-Parametric Statistical Analysis was also used in this work for the statistical processing of the experimental results [1].

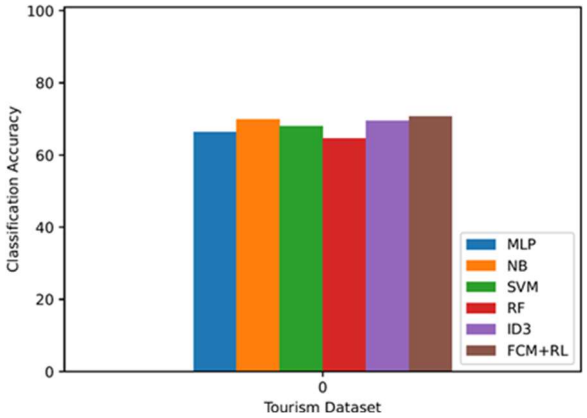


Fig. 2. Behavior of a reinforcement learning problem.

Experiment 1: Compare the method proposed (FCM+CWW+RL) with 3 algorithms: MLP, NB and ID3 which are implemented in WEKA tool.

Goals: Determine if the accuracy of the propose method FCM+CWW+RL is significantly superior to accuracy of MLP, ID3 and NB. In tables 2 and 3 are shown these results. Table 2 shows the list of all datasets used to compare the accuracy of this proposal with a well-known classification algorithm.

In table 3 it can be observed the results of applying the Holm test, and how the null hypothesis is rejected for all the algorithms with which this proposal is compared to.

5 Real Case Study

In tourism, we can find many challenges, but there is one that represents a continuous race to success. This challenge is to build the best network of stake- holders to keep services continuously at the same level. Suppliers vary according to the needs of the tourism facility (proximity, better products, fresh fish, fresh vegetables, etc.).

In the city of Camagey, Cuba, all hotels have been trying for a long time to find the best suppliers for their services. From an analysis of data and scenarios about these processes, we decided to collect data from suppliers and hotels to build a FCM and test the method proposed in this paper. Figure 2 details the results of the method proposed in this research using stakeholder and hotel data. As can be seen, the accuracy of our method outperforms the rest of the known algorithms selected for comparison.

6 Conclusions

In this paper, we have studied a new learning method that improves the classification in the FCM where the activation values of the concepts and the causal weights are described by means of linguistic terms. The results obtained show the effectiveness of using RL as a learning method to adjust the relations matrix in the FCM. According to the results presented in this paper, the approach is statistically better than MLP, NB, SVM, RF and ID3 in in terms of classification accuracy. Further research may include

performing hybridization with the method proposed in this research and that of [4] although many variations are possible.

References

1. Alcala-Fernández, J., Fernández, A., Luengo, J., Derrac, J.: Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing, vol. 17, no. 2, pp. 255–287 (2010)
2. Batista, G. E. A. P. A., Prati, R., Monard, M. C.: A study of the behaviour of several methods for balancing machine learning training data. Association for Computing Machinery, vol. 6, no. 1, 20–29 (2004) https://doi.org/10.1145/1007730.1007735
3. Axelrod, R.: Structure of decision: The cognitive maps of political elites. Princeton University Press (1976)
4. Balmaseda, F., Filiberto, Y., Frias, M., Bello, R.: A new approach to improve learning in fuzzy cognitive maps using reinforcement learning. Applied Computer Sciences in Engineering, pp. 226–234 (2019) doi: 10.1007/978-3-030-31019-6_20
5. Chen, Y., Mazlack, L., Lu, L.: Learning fuzzy cognitive maps from data by ant colony optimization. In: Proceedings of the 14th annual conference on Genetic and evolutionary computation, pp. 9–16 (2012) doi: 10.1145/2330163.2330166
6. Delgado, M., Verdegay, J. L., Vila, M. A.: On aggregation operations of linguistic labels. International Journal Intelligent Systems, vol. 8, no. 3, pp. 351–370 (1993) doi: 10.1002/int.4550080303
7. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, vol. 7, pp. 1–30 (2006)
8. Dickerson, J. A., Kosko, B.: Virtual worlds as fuzzy cognitive maps. Presence Teleoperators and Virtual Environments, vol. 3, no. 2, pp. 173–189 (1994) doi: 10.11 62/pres.1994.3.2.173
9. Dubois, D., Prade, H.: Fuzzy sets and systems: Theory and applications. Academic Press (1980)
10. Filiberto, Y., Bello, R., Nowe, A.: A new method for personnel selection based on ranking aggregation using a reinforcement learning approach. Computaci´on y Sistemas, vol. 22, no. 2 (2018) doi: 10.13053/CyS-22-2-2353
11. Frias, M., Filiberto, Y., N´apoles, G., Garc´ıa-Socarr´as, Y., Vanhoof, K., Bello, R.: Fuzzy cognitive maps reasoning with words based on triangular fuzzy numbers. Advances in Soft Computing, pp. 197–207 (2018) doi: 10.1007/978-3-030-02837-4_16
12. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences, vol. 180, no.10, pp. 2044–2064 (2010) doi: 10.1016/j.ins.2009.12.010
13. García, S., Herrera, F.: Evolutionary under-sampling for classification with imbalanced data sets: Proposals and taxonomy. Evolutionary Computation, vol. 17, no. 3, pp. 275–306 (2009) doi: 10.1162/evco.2009.17.3.275
14. Herrera, F., García, S.: An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research, vol. 9, no. 89, pp. 2677–2694 (2008)
15. Herrera, F., Martínez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems, vol. 8, no. 6, pp. 746–752 (2000) doi: 10.1109/91.890332
16. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70 (1979)
17. Vázquez-Huerga, A.: A balanced differential learning algorithm in fuzzy cognitive maps. In: 16th International Workshop on Qualitative Reasoning (2002)

18. Iman, R., Davenport, J.: Approximations of the critical region of the Friedman statistic. Communications in Statistics-Theory Methods, vol. 9, no. 6, pp. 571–595 (1980) doi: 10.1080/03610928008827904
19. Kosko, B.: Fuzzy cognitive maps. International Journal of Man, Machine Studies, vol. 24, no. 4, pp. 65–75 (1986) doi: 10.1016/S0020-7373(86)80040-2
20. Koulouriotis, D. E., Diakoulakis, I. E., Emiris, D. M.: Learning fuzzy cognitive maps using evolution strategies: A novel schema for modeling and simulating high-level behavior. In: Proceedings of the 2001 Congress on Evolutionary Computation, vol. 1, pp. 364–371 (2001) doi: 10.1109/CEC.2001.934413
21. Nápoles-Ruiz, G. R.: Algoritmo para mejorar la convergencia en mapas cognitivos difusos sigmoidales. Doctoral dissertation, Universidad Central "Marta Abreu" de Las Villas, Facultad de Matemática, Física y Computación, Departamento Ciencias de la Computación (2014)
22. Narendra, K., Thathachar, M.: Learning automata: An introduction. Prentice-Hall International (1989)
23. Papageorgiou, E. I.: Learning algorithms for fuzzy cognitive maps– a review study. IEEE Transactions on Systems, Man, and Cybernetics, vol. 42, no. 2, pp. 150–163 (2012) doi: 10.1109/TSMCC.2011.2138694
24. Parsopoulos, K. E., Papageorgiou, E. I., Groumpos, P., Vrahatis, M. N.: A first study of fuzzy cognitive maps learning using particle swarm optimization. In: IEEE Congress on Evolutionary Computation, vol. 2, pp. 1440–1447 (2003) doi: 10.1109/ CEC.2003.1299840
25. Sheskin, D. J.: Handbook of parametric and nonparametric statistical procedures. Chapman and Hall (2003)
26. Stach, W., Kurgan, L., Pedrycz, W., Reformat, M.: Genetic learning of fuzzy cognitive maps. Fuzzy Sets and Systems, vol. 153, no. 3, pp. 371–401 (2005) doi: 10.1016/j.fss.2005.01.009
27. Sutton, R. S., Barto, A. G.: Reinforcement learning: An introduction. The MIT Press (2017)
28. Thathachar, M., Sastry, P. S.: Networks of learninga: Techniques for online stochastic optimization. Springer Science+Business Media, LLC (2004) doi: 10.1007/978-1-4419-9052-5
29. Wauters, T., Verbeeck, K. Causmaecker, P., Vanden-Berghe, G.: Fast permutation learning. Lectures Notes on Computer Science, vol. 7219, pp. 292–306 (2012) doi: 10.1007/978-3-642-34413-8_21
30. Zadeh, L. A.: Outline of a new approach to the analysis of complex systems ad decision processes. IEEE Transactions Systems, Man, and Cybernetics, vol. SMC-3, no. 1, pp. 28–44 (1973) doi: 10.1109/TSMC.1973.5408575

ISSN 1870-4069

Diabetic Retinopathy Detectio Via Local Binary Patterns

David Ferreiro-Piñeiro, Ivan Olmos Pineda,
Arturo Olvera López

Benemérita Universidad Autónoma de Puebla,
Mexico

david.ferreiro@alumno, {ivan.olmos,
jose.olvera}@correo.buap.mx

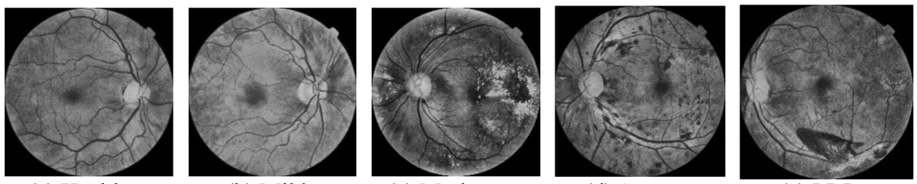
Abstract. Diabetic Retinopathy (DR) is the leading cause of preventable blindness for working-age people. Clinical detection methods are expensive, time-consuming, and subjective. Therefore, automated techniques have been proposed to develop complementary objective tools to support the diagnosis that facilitate its detection. In this work, we offer an approach for detecting DR in retinal images using a texture descriptor. The behavior of two different implementations of the Local Binary Patterns (LBP) is analyzed in the APTOS database, using sliding windows mechanisms to extract a more significant amount of detail from the descriptors used. The classification is carried out via a Support Vector Machines (SVM) whose training and evaluation are performed using cross-validation. The experimental results of the two implementations are compared; the model based on uniform LBP obtained the best performance with an accuracy of 0.935, a sensibility of 0.924, and a specificity of 0.947, demonstrating that the extraction of relevant features using uniform LBP guarantees the detection of DR.

Keywords: Diabetics retinopathy, LBP, uniform LBP.

1 Introduction

The lifestyle habits in populations determine the evolution of new diseases such as Diabetes. This kind of disease causes an increment in sugar levels in the blood, affecting retinal tissue and evolving into a condition known as Diabetic Retinopathy (DR). According to the World Health Organization (WHO), 146 million people suffer from DR, which is the leading cause of blindness for productive people [1, 2].

DR is a progressive and degenerative disease that causes pathological changes in retinal tissue, and its diagnosis depends on the detection of lesions through observation and analysis from ophthalmologists. According to the lesions found, ophthalmology classifies the disease in five stages: healthy, mild, moderate, severe, and proliferative (PDR) (Fig. 1 shows the evolution of these kinds of lesions).



(a) Healthy (b) Mild (c) Moderate (d) Severe (e) PDR

Fig. 1. Clinic evolution to the Diabetic Retinopathy (Source: Decencière et al. [5]).

The clinical manifestations of the disease are microaneurysms, hemorrhages, exudates [3], and new defective blood vessel or vascular abnormalities, which can lead to retinal detachment and vision loss.

The clinic method for diagnosis is expensive and requires qualified specialists; the diagnosis quality sometimes depends on subjective factors such as exhaustion, stress, and experience. For these reasons, the WHO recommends strengthening research to expand technological advances to help people with this type of condition [4].

Timely detection of DR has become a field of research for the development of computer vision and artificial intelligence applications. The main approaches are oriented to the detection of specific lesions to conduct classification tasks or to the development of models based on deep learning.

The proper diagnosis depends not only on the detection of the lesions but also on their location in the retinal tissue. In the present investigation, a diagnostic model for DR is proposed using classic computer vision techniques that allow the detection and identification of specific lesions to be dispensed with and guarantee the diagnosis of the disease; considering the limited availability of retinal images, it does not suggest using deep learning approaches.

Diagnosis determines the presence of the disease in retinal images. The proposed models are helpful as a complementary diagnostic mechanism for modern technologies that allow the attention of ophthalmological care while reducing cost and guaranteeing access to the most vulnerable population.

This work is organized as follows: Section 2 developed a study of state of art related to the use of SVM in the detection of DR. Section 3 presents a methodological proposal from the pre-processing of the images, the study of the different implementations of the LBP, the experimental development, and the discussion of the results. Finally, the conclusions of the work are presented.

2 Related Works

The analysis of the literature allowed us to identify five main working approaches during the automatic diagnosis of DR: based on regions (oriented to the detection of specific lesions designing the relevant characteristics), modification of known Convolutional Neural Network (CNN) models, hybrids (use CNN to extract relevant features, for the classification they use other traditional classifiers), development of new architectures and hyperparameter tuning (improve system performance by modifying specific parameters).

The work proposed by AbdelMaksoud et al. [6] is oriented to the detection and analysis of the pathological changes that affect the retina during the development of DR, performing the segmentation of specific lesions (microaneurysms, exudates, and hemorrhages). Characteristics are extracted using different techniques such as the Gray Level Co-Occurrence Matrix, the area of the blood vessels, the bifurcation points, and determined the number of lesions in the images.

After the characteristics are extracted, a classification model is developed based on multilabel SVM on the public databases DRIVE, Messidor, Stare, and IDRiD, reporting an accuracy of 0.892. Issac et al. [7] perform a quantitative analysis of the red, bright lesions and the optic disc present in the images. From these data extracted, ten characteristics and the classification using the Messidor-1 and DIARETDB0 databases.

The authors highlight lesions by intensity normalization and thresholding for detection to improve experimental results. They obtained an accuracy of 0.9213, a sensitivity of 0.9285, and a specificity of 0.8 in the DIARETDB0 database. Another approach (Katada et al.) [8] proposes a hybrid architecture of Inception-V3 and SVM on a private database of two hundred images corresponding to patients of Japanese ethnicity and on the EyePACS database.

The main contribution is that the models can be abstracted from ethnic traits, representing an advantage due to the limited availability of data from non-Anglo-Saxon patients. A sensitivity of 0.815 and 0.908, and specificity of 0.719 and 0.9, are reported for EyePACS and the private database, respectively. Another approach estimates the probability of disease development from the location of microaneurysms and hemorrhages on retinal images [9].

The idea developed by the authors not only segment the lesions but also find them and determine their relevance. The authors employ transfer learning techniques to mitigate limited data availability. Son et al. [10] develop independent models to detect twelve different lesions: hemorrhages, hard exudates, soft exudates, retinal scars, any vascular abnormality, and nerve fiber layer defects, among others; these algorithms are based on the information provided by the region.

The proposed architecture allows detecting multiple lesions, combining various models, and visualizing them with a heat map as an auxiliary output to identify the aspects considered by the algorithm to generate the final classification. There is a tendency to use approaches based on deep learning; these approaches require a considerable volume of data to perform their training and validation.

In addition, recent works are oriented to supply an adequate solution to the problem and explain how that solution is reached. This allows the introduction of external specialists to validate the proper functioning of the systems.

3 Methodological Proposal

The methodological proposal is shown in Fig. 2. The first step is a pre-processing phase; then, relevant features are extracted to perform the classification process. This article is oriented to the analysis of the texture in the different regions of the image to conduct the diagnosis of DR without it being necessary to detect the specific lesions present in the tissue. The proposal classifies the input image as healthy or sick, although it can be

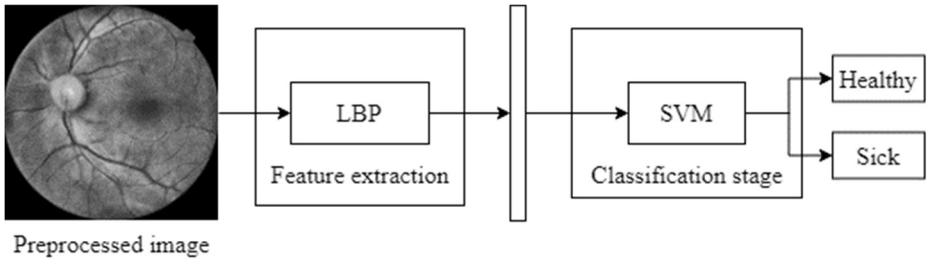


Fig. 2. Proposed methodological scheme.

extended to the detection of PDR, which considers the degree of evolution of the disease.

3.1 Pre-Processing Steps

The models were developed and evaluated on images from a public tagged repository managed within the framework of the 4th Symposium of the Asia Pacific Tele-Ophthalmology Society (APTOS). These images were captured under different lighting conditions, techniques and instruments and included the retinal tissue, its vascular network, macula, and optic disc. This region is surrounded by an area that does not provide helpful information, making processing work difficult.

The pre-processing, the retinal tissue is delimited by segmenting the brightness channel by applying the Otsu method in the HSV color space, determining the transition points between the tissue and the surrounding area that define the proper working area. Subsequently, a Gaussian filter is applied with $\sigma = 1$, which reduces the Gaussian noise; a change is made in the color space using the LAB model, which allows the separation of the lighting treatment on the chroma channels.

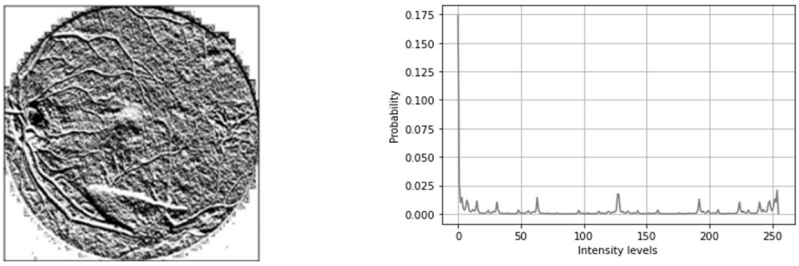
Contrast-Limited Adaptative Histogram Equalization (CLAHE) is applied on the L channel, modifying the image's lighting without affecting the color components, avoiding the introduction of artifacts that can be confused with lesions. Finally, the dimensions of the pictures are homogenized to 600 × 600 pixels.

3.2 Feature Extraction

Feature extraction is a crucial stage for the development of a classification model. It allows for defining the distinctive features for detecting the disease and describing its clinical grade. This work proposes to develop this process considering the textures of retinal images. The textures are repetitive patterns in local intensity differences that allow differentiated structures [11].

Different texture descriptors include Gray Level Co-Occurrence Matrix (GLCM), Binary Local Patterns (LBP), and Gabor filters, among others. This descriptor aims to quantify the qualities that we intuitively define as rough, silky, or soft and that we commonly use depending on the intensity variation of the pixels.

The main differences between these descriptors are how they process spatial information. GLCM generates second-order statistics from the definition of co-occurrence matrices, according to the distances and the direction of analysis.



(a) LBP image (b) LBP image histogram

Fig. 3. Analysis of the behavior of the classic LBP extractor.

This descriptor suffers from high dimensionality, consumes large amounts of memory, and its derived features have a high correlation [12].

Gabor filters emulate the human vision process but require adjustments regarding the orientation of the filters. Orientation is a critical aspect because the detection of the characteristics in the images depends on their excellent tuning. Filter banks are used to solve this problem by processing the images in different orientations, and this increases the computational requirements.

The LBP approach is considered to extract texture features using a circular neighborhood where the texture is defined as the joint distribution of the gray levels described as the equidistance between the central pixel and the radius of the neighborhood. From the fundamental characteristics of the previously exposed descriptors, it was decided to use the LBP during the experimental development of the present investigation. According to [13], the LBP descriptor can be expressed as:

$$\text{LBP} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \tag{1}$$

$$s(g_p - g_c) = \begin{cases} 1 & \text{if } g_p - g_c \geq 0, \\ 0 & \text{if } g_p - g_c < 0, \end{cases} \tag{2}$$

where g_c is the central pixel intensity, g_p is the intensity of the neighboring pixel and finally 2^p quantization weights. When applying this descriptor, the exudates are represented as craters on the retina's surface; the hemorrhages or microaneurysms are described as plains or elevations in the retinal tissue; these results can be visualized in Fig. 3a, the relief generated by the hemorrhage is observed, allowing its identification.

LBP allows obtaining the histograms that describe the image and is invariant to rotation. The property of invariance to rotation is crucial, the retinal images can be taken from different angles, and descriptors must characterize the presence of the disease regardless of the spatial distribution of the characteristic lesions in the retina.

Then, this descriptor allows us to differentiate between the types of the characteristic lesion and, therefore, could be used as a discriminator. Fig. 3-b shows the input image's histogram; it can be noticed how the image is distributed over the entire dynamic range.

Histogram-based approaches can be ambiguous in describing large regions; sliding window mechanisms are used to avoid this problem. The descriptor for specific areas

Table 1. Comparison considering some state-of-the-art methods.

Model	Auc	Acc	Se	Sp
Uniform LBP	0.97	0.935	0.924	0.947
Classical LBP	0.97	0.927	0.924	0.93
AbdelMaksoud et al. [6]	0.89		0.85	0.85
Issac et al. [7]		0.92	0.92	0.8
Katada et al. [8]			0.90	0.80

is determined and concatenated to form a single feature vector. This approach strengthens feature detection and, when using LBP, helps guarantee the descriptor's translation invariance. To use this approach, the proper selection of the size of the windows is essential; large windows lose the details, and small windows cannot determine the presence of specific lesions.

The output feature vector would be equal to $\vec{V} = 256 \times N$, where N is the number of sliding windows and 256 are the intensity levels of the histograms using the proposed approach. This approach consumes expensive computing resources.

One way to reduce computational requirements is to consider the presence of uniform patterns during LBP encoding. When the pixels surrounding the central pixel do not alternate or contain at most two transitions, this pattern is considered uniform, otherwise called non-uniform [14]. This descriptor can be expressed as:

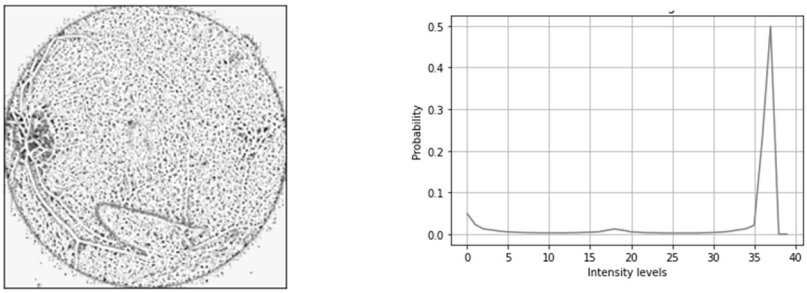
$$\mathrm{LBP}_{P,R}^{\mathrm{riu2}} = \begin{cases} \sum_{P=0}^{P-1} s(g_p - g_c) & \text{if } U(LBP_{P,R} \leq 2), \\ P+1 & \text{other case,} \end{cases} \tag{3}$$

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c)| + \sum_{P=1}^{P-1} |s(g_P - g_c) - s(g_{P-1} - g_c)|, \tag{4}$$

where P number of pixels in the circular neighborhood determines the angular resolution, R radius of the circular neighborhood determines the spatial resolution, and finally $LBP_{P,R}$ is a measure of uniformity. It has shown that combinations of uniform patterns and the rotation invariance property can be encoded with $P + 1$ levels of intensity, and all non-uniform patterns are represented with an additional intensity level.

It might be thought that a significant loss of information may occur by using uniform LBP. However, based on the uniform LBP results (**¡Error! No se encuentra el origen de la referencia.**-a), it can be noticed that the details that define the presence and location of a lesion are still preserved. According to these results, it is possible to represent relevant information considering a lower amount of intensity level.

In **¡Error! No se encuentra el origen de la referencia.**-b, the histogram generated using the uniform approach $\mathrm{LBP}_{36,3}^{\mathrm{riu2}}$ is depicted. In this case, the output vector is modified and will have a length of $\vec{V} = 38 \times N$ due to the properties of the analyzed images, the descriptor considers an angular resolution of 10°.



(a) Uniform LBP image. (b) Uniform LBP image histogram.

Fig. 4. Uniform LBP image analysis.

3.3 Classification Stage

Feature extraction was performed considering the use of sliding window techniques. The size of the windows was defined in 100 x 100 pixels, with a 50% overlap between adjacent windows to increase the redundancy of the extracted features.

A total of 144 windows or analysis segments were generated in which the LBP vector is determined. The uniform LBP vector will have a dimension of 5472 elements as opposed to the 36720 elements of the classical LBP vector.

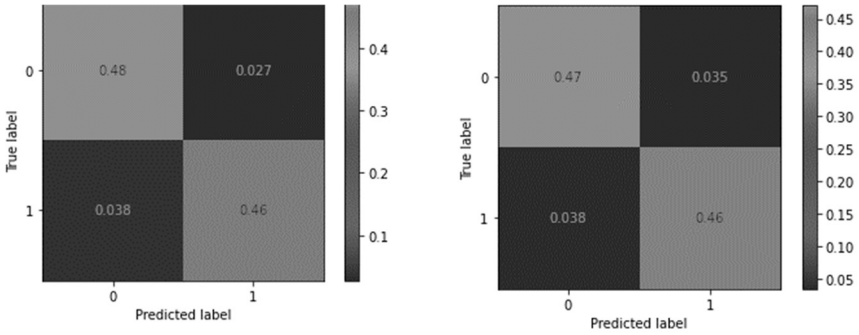
The APTOS database contains a population of 5593 images, of which 3662 are labeled [15]. These images were classified according to the degree of clinical evolution of the disease (healthy, mild, moderate, severe, proliferative). For the present work, they were grouped in a binary way into healthy or sick, and stratified random sampling to preserve the ratio classes in the original training set is conducted. The training and test sets were formed considering 80% and 20%, respectively.

In addition, the selection of the hyperparameters of the models was conducted using a genetic optimization algorithm. This algorithm was applied to the models developed in the classical LBP and uniform LBP features. An SVM with a degree three polynomial kernel was designed to address the uniform LBP features, and an SVM with a radial basis function kernel for de classical LBP features.

The model was validated, applied a ten cross-validation, and analyzed the area under the ROC curve as a metric of interest, which is a standard metric for comparing the experimental results of applications oriented to medical services [16]. In Fig. 5, the confusion matrices generated when evaluating the classification models on the test images are depicted.

The metric used to evaluate the performance of the developed models are the area under the ROC curve (Auc), accuracy (Acc), sensitivity (Se), and specificity (Sp). A priori analysis allows us to establish that the model based on uniform LBP has a better behavior than the classic LBP model. However, in medical environments, it is more significant to achieve high sensitivity, reducing the occurrence of false negatives. False positives can be corrected by performing complementary clinical tests, unlike false negatives that could stop the treatment of the disease.

The uniform LBP model has a higher sensitivity by reporting fewer false negatives, around 3.8% versus 2.7% false positives. While the model developed on classic LBP



(a) Uniform LBP (b) Classical LBP

Fig. 5. Experimental confusion matrices.

has the opposite behavior, it has a higher number of false negatives (3.8%) than the false positives detected. Therefore, both models have the same sensitivity, but the uniform LBP has better specificity and accuracy. In Table 1, the main experimental results obtained are observed in comparison with other models reported in state-of-the-art. Based on our preliminary results, the proposed methodology allows the detection of characteristic lesions simply in the form of reliefs in the retinal tissue.

It, therefore, allows detecting the presence of the DR. The uniform LBP descriptor guarantees to extract the relevant features to conduct the detection using a vector of a small length without losing relevant information, which is particularly useful in limited computational resources environments.

4 Conclusions

Two classification models were developed and evaluated using support vector machines and texture descriptors to detect Diabetic Retinopathy in retinal pictures. The descriptors used guaranteed the exact spatial and angular resolution but differed in their form to encode the results. The implementations differed in the number of elements considered in the feature vectors.

The model developed on the uniform LBP obtained a similar performance in the ability to detect the presence of the disease (sensitivity) and in the area under the ROC curve and superior performance in the specificity and accuracy of the classification when compared with the model developed on the classic LBP.

From this, it was found that the number of features used does not determine, in general, the quality of the classification models. The model based on the classic LBP is sensitive to the presence of artifacts or other variations in the images, aspects to which the uniform implementation is invariant.

In addition, the proposed model of better results was compared with some works reported in state of the art, with encouraging results, although it would be interesting to evaluate this model on the same databases used by other authors. In future work, it is necessary to evaluate the behavior of the model developed on other public databases to validate the results and be able to make an objective comparison of the results.

References

1. World Health Organization: World report on vision. World Health Organization (2019) www.who.int/publications/i/item/9789241516570
2. Kumar, S., Adarsh, A., Kumar, B., Singh, A. K.: An automated early diabetic retinopathy detection through improved blood vessel and optic disc segmentation. Optics and Laser Technology, vol. 121, pp. 105815 (2020) doi: 10.1016/j.optlastec.2019.105815
3. Saranya, P., Prabakaran, S.: Automatic detection of non-proliferative diabetic retinopathy in retinal fundus images using convolution neural network. Journal of Ambient Intelligence and Humanized Computing, Springer Science and Business Media (2020) doi: 10.1007/s12652-020-02518-6
4. World health organization: Recommendations on digital interventions for health system strengthening. WHO Guideline (2019) www.who.int/publications/i/item/9789241550505
5. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J. C.: Feedback on a publicly distributed image database: the messidor database. Image Analysis and Stereology, Slovenian Society for Stereology and Quantitative Image Analysis, vol. 33, no. 3, pp. 231 (2014) doi: 10.5566/ias.1155
6. AbdelMaksoud, E., Barakat, S., Elmogy, M.: A comprehensive diagnosis system for early signs and different diabetic retinopathy grades using fundus retinal images based on pathological changes detection. Computers in Biology and Medicine, vol. 126, pp. 104039 (2020) doi: 10.1016/j.compbiomed.2020.104039
7. Issac, A., Dutta, M. K., Travieso, C. M.: Automatic computer vision-based detection and quantitative analysis of indicative parameters for grading of diabetic retinopathy. Neural Computing and Applications, vol. 32, no. 20, pp. 15687–15697 (2018) doi: 10.1007/s00521-018-3443-z
8. Katada, Y., Ozawa, N., Masayoshi, K., Ofuji, Y., Tsubota, K., Kurihara, T.: Automatic screening for diabetic retinopathy in interracial fundus images using artificial intelligence. Intelligence-Based Medicine, vol. 3–4, pp. 100024 (2020) doi: 10.1016/j.ibmed.2020.100024
9. Zago, G. T., Andreão, R. V., Dorizzi, B., Teatini Salles, E.: Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Computers in Biology and Medicine, vol. 116, pp. 103537 (2020) doi: 10.1016/j.compbiomed.2019.103537
10. Son, J., Shin, J. Y., Kim, H. D., Jung, K. H., Park, K. H., Park, S. J.: Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. Ophthalmology, vol. 127, no. 1, pp. 85–94 (2020) doi: 10.1016/j.ophtha.2019.05.029
11. Chaki, J., Dey, N.: Texture feature extraction techniques for image recognition. Springer Singapore (2020) doi: 10.1007/978-981-15-0853-0
12. Hall-Beyer, M.: Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales. International Journal of Remote Sensing, Informa UK Limited, vol. 38, no. 5, pp. 1312–1338 (2017) doi: 10.1080/01431161.2016.1278314
13. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987 (2002) doi: 10.1109/tpami.2002.1017623
14. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037–2041 (2006) doi: 10.1109/tpami.2006.244
15. Kagle: APTOS 2019 Blindness Detection (2022) www.kaggle.com/competitions/aptos2019-blindness-detection/overview

16. Hanley, J. A., McNeil, B. J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, Radiological Society of North America, vol. 143, no. 1, pp. 29–36 (1982) doi: 10.1148/radiology.143.1.7063747

ISSN 1870-4069

Sign Language Recognition through Manual and Non-Manual Features

Daniel Sánchez-Ruiz, José Arturo Olvera-López,
Iván Olmos-Pineda

Benemérita Universidad de Puebla,
Facultad de Ciencias de la Computación,
Mexico

daniel.sanchezruiz@viep.com.mx, jose.olvera@correo.buap.mx,
iolmos@cs.buap.mx

Abstract. Deaf community uses sign language as its main form of communication; however, most of the speaking community does not know how to understand that language, therefore the sign language recognition through technological developments has been an area of great interest for years. In this work, a proposal for this problem is presented, where regions of interest detection, manual and non-manual features extraction are carried out and for the recognition some BiLSTM networks with different architectures are used. The results obtained are an 73.99% accuracy, which are promising for the upcoming experiments. Finally, various actions are presented with the aim of improving the results as future work.

Keywords: Sign language recognition, computer vision, pattern recognition.

1 Introduction

People are considered to have a hearing loss when they are not able to hear under a hearing threshold of 25dB or less in both ears. Around 430 million people worldwide have disabling hearing loss, and it is estimated that by 2050 over 700 million people will suffer this kind of disability [1].

Hearing loss is one of the most common chronic impairments that appear with age as degeneration of sensory cells. It results from different congenital or acquired causes (e.g.: genetic causes, complications at birth, infectious diseases, exposure to excessive noise, among others).

Sign languages are classified as natural languages [2], which are used by the deaf community as their principal way of communication. Sign languages' visual, spatial nature and their variability, present a considerable research problem to be solved through technological developments.

Numerous areas are involved, such as linguistics, medicine, machine learning, computer vision, natural language processing, and computer graphics. Sign Language Recognition (SLR) is the scientific area responsible for capturing and translating sign speech using computer vision and artificial intelligence techniques [3]. Considering the

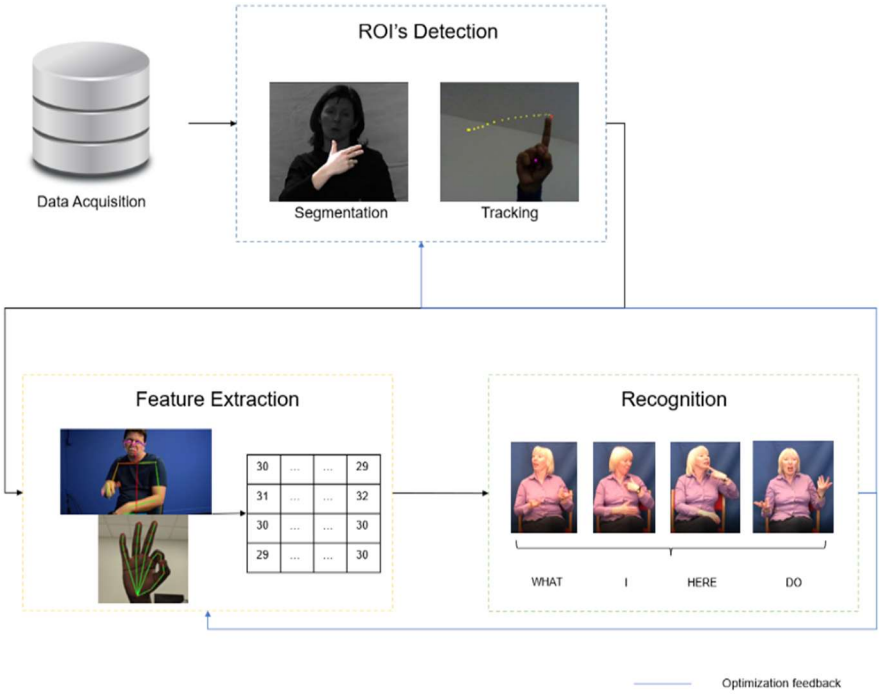


Fig. 1. Diagram of the proposed methodology for sign language recognition.

importance of sign languages for the communication of millions of people across the world and the rapid technological developments, this work proposes a methodology for sign language recognition employing several features.

The main proposal regarding the related work is to extract non-manual features based on the estimation of gaze and head pose along other well studied features; these two descriptors have not been studied thoughtfully, as can be seen in some works [4]. Another difference with recent works is the region of interest detections unlike the use of deep learning techniques as feature extractor, where is common to discard it, this is done to avoid adding unnecessary noise in the recognition phase.

This paper is organized as follows: in section 2 related works are analyzed and presented; section 3 is focused on the proposed methodology; section 4 describes the experiments that were designed and the obtained results and finally, in section 5 conclusions and future work are listed.

2 Related Work

SLR has been a research area very active since 90's [5], but recently important advances have been reported. At the beginning most of the studies focused in the used of gloves or haptic sensors to segment and track hands [6–8]. Nonetheless, deaf community felt very intrusive these types of methods because it can create practical difficulties in daily life and often limit their movements. By all these reasons, recent works are mainly

Fig. 2. Frames of example from the LIBRAS dataset.

focused on solutions based on computer vision, where the only necessary equipment are cameras.

As in any pattern recognition system, features extraction is an important stage. In SLR systems using cameras as input capture devices depend on the computer vision and image processing techniques. Common tasks performed are hand shape estimation, gesture segmentation, contour and boundary modeling, or color and motion cue identification.

Sign Languages have two types of features: manual and non-manual. Manual features consist in spatial and temporal descriptors base in the hand region; shape, position and motion are the most employed. As its name suggests non-manual features are all the cues related with the rest of the body.

They contain relevant information, which helps to recognize sign gestures with better accuracy. Non-manual features convey semantic or lexical properties, but also syntactical and grammatical functions, e.g.: negation, clausal type, question status, topics, or emphasis.

Several non-manual features have been studied, the principal are facial expressions and body pose. The SLR research can be classified in two principal types of investigation: isolated (ISLR) and continuous sign language recognition (CSLR). ISLR involves the recognition of a letter or a word at a time.

Some ISLR approaches employ Leap Motion (LM) sensors to recognize isolated words [9, 10]. In the former approach, they use fingertip information and their correlation, then a Support Vector Machine (SVM) [28] is used as recognition method to get an accuracy of 91.28% for ten ASL digits, whereas in the latter approach, they used extracts of 3D information and again a SVM. This system shows the best accuracy of 96.50% for ten ASL digits.

Kumar et al. [11] recognized 50 isolated signs using Kinect and LM sensors, an accuracy of 40.23% for all sign gestures are the results obtained. The principal disadvantage is that only manual features are considered. Ibrahim et al. [12] recognized 30 isolated Arabic signs and obtained an accuracy of 97%. However, to be implemented in real-time continuous sign sentences, more experiments in bigger vocabulary needs to be addressed.

CSLR concerns in recognizing one or more complete sentences. CSLR is more challenging than ISRL; problems of occlusion, alignment, or sign gestures identification in respect of transition movements are some of the difficulties that need to be considered.

It is difficult to recognize the transition movements because they are very subtle between all the different signs in data for CSLR, for this reason it is a topic of relevant interest in the research community. Common approaches to solve or mitigate this problem is eliminate epentheses movements (transition movements) by explicit modeling, implicit modeling or simply ignoring the transition movements.

Kong and Ranganath [13] presented a probabilistic approach based on the design and recognition of sign sub-segments and produced an 81.6% accuracy, the drawback is that movement epentheses are labeled manually. Li et al. [14] presented a scalable approach ignoring transition movements, the proposal gives an accuracy of around 87%. The inconveniences of the proposed approach are the use of a small vocabulary and its computer's execution time, which is considerable.

Elakkiya and Selvamani [15] proposed an automatic sign language classification, where they break down signs into subunits without any prior knowledge about the gestures. A Bayesian parallel hidden Markov model is used, its function is to combine manual and non-manual subunit features, but besides that it also handles the problem of movement ambiguities. An 82.1% of accuracy with signer independence was obtained as a result.

3 Methodology

The stages of the proposed methodology are depicted on the diagram in Figure 1. The first step consists of obtaining the input data to be used, the second and third steps are related to region of interest (ROI) detection and feature extraction, respectively; finally, the recognition task occurs. In the following subsections each one of these steps are described in detail.

3.1 Data Acquisition

LIBRAS (Brazilian Sign Language) dataset [16] was used for the experimental part, in particular Florianópolis' data, which contains 639 records, all the videos have a resolution of 640x414 pixels, with a refresh rate of 30 frames per second; besides that, an EAF file for the annotation of the signs is incorporated to each record. The topics that are covered in the dataset are dates, fruits, numbers, literature, interviews among others.

In every record two persons are present in the room having a conversation, which makes this dataset of continuous type. Four videos were recorded, one with an aerial angle, one with a lateral view of both persons and the last two with a direct view of each person.

For the problem the latter are the most suitable to use, however, as this dataset was not thought for SLR systems, in some records only one of the persons is gesticulating signs while the other is only watching, for this reason some videos are discarded. Figure 2 shows a couple of frames from one of the videos as an example.

3.2 ROI's detection

Based on sign language grammar [5], the regions of interest that were defined are the hands in order to extract manual descriptors and the body and the face in order to extract

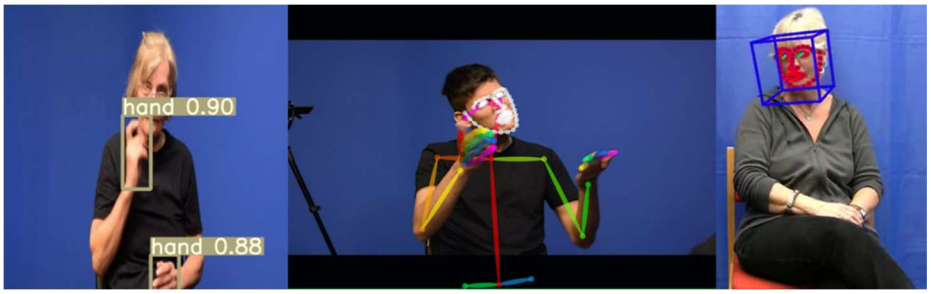


Fig. 3. Examples of ROI's detection through YOLOv5, OpenPose and OpenFace.

non-manual descriptors. For the task of detecting the hands' ROI, YOLOv5 system [17] was used, which is the state of the art in the object detection task.

However, since the features that are present in the data are very specific (deformations and occlusions), a training process from scratch with a manual annotated subset of LIBRAS images is performed to obtain a custom model.

For the body posture estimation OpenPose was used [18], which in addition to estimating the key points referring to the body joints, also brings the possibility of estimating key points related to the regions of the hand and face. Additionally, OpenFace [19] was also used for obtaining more characteristics related to the face and head. In the Figure 3 some examples of the obtained results of this stage are shown.

3.3 Feature Extraction

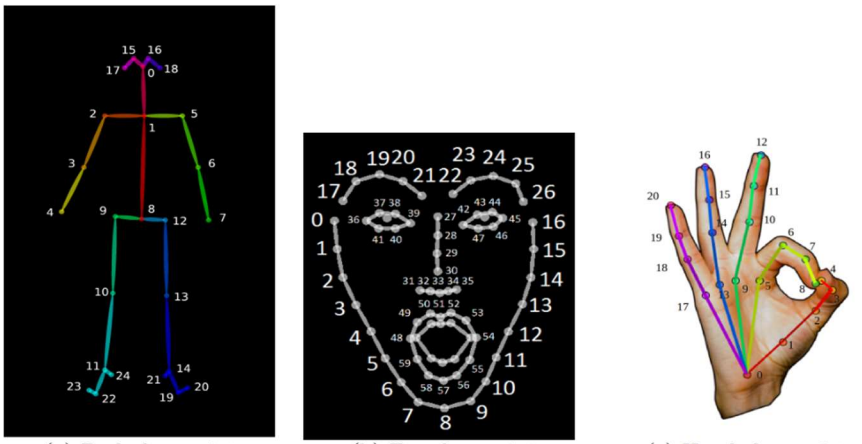
Aloysius and Geetha [20] stated that with approaches based on deep learning, such as the use of convolutional neural networks as feature extractors, it is no longer necessary to do ROIs detection and feature extraction locally.

Although the results have improved considerably, in most of these works in the input images irrelevant information is not previously discarded (context or even parts of the body such as legs that are not necessary). Furthermore, deep learning-based approaches work best with large amounts of data, which is not the case in most of the existing datasets.

For these reasons, the detection of regions of interest and the local extraction of descriptors based on manual and non-manual characteristics were proposed. As Koller [4] describes, several works have studied the relevance in the use of descriptors based on manual features (hands).

However, non-manual features (body and head) are also important in sign languages, and as Koller showed, they have been less explored, those related to the position of the head, or the direction of gaze have not been explored to the best of the knowledge of the authors. Taking this into account, the following features are extracted.

- Coordinates (x,y) for each hand. This relative to the centroid of the envelope frame detected with YOLOv5.
- Approximate speed for each hand. Tracking the change of the centroids' position every two distinct intervals of time (every 3 frames).



(a) Body key points (b) Face key points (c) Hands key points

Fig. 4. Key points obtained through OpenPose.

- Euclidean distance between selected key points related with facial expressions and hands. The OpenPose points considered are (0-9), (0-2), (17-20), (13-16), (9-12), (5-8), (2-4), (51,57), (48-54), (33-51), (19-37) and (24-44); they can be visualized in Figure 4a and Figure 4c.
- Key points coordinates (x, y) related to the arms. The points considered are (3, 4, 6, 7) and can also be seen in Figure 4b.
- Angular coordinates (x, y) of the gaze direction. Coordinates in radians averaged for both eyes and obtained with OpenFace.
- Rotation in radians around the X, Y, Z axes. Values obtained through OpenFace, which provides the head posture.

3.4 Sign Language Recognition

Since the problem to solve is a sequential problem, the recognition method to be used was a BiLSTM network, which have proven to be useful in several works [4, 20, 3]. Three architectures are proposed, the first one serves as a base to explain the last two and is shown in Figure 5, it is composed of an LSTM layer that is bidirectional, a fully connected layer and a softmax layer.

The second architecture has the BiLSTM layer, followed by a dropout layer, a ReLU layer, a fully connected layer and the softmax layer. Finally, the last architecture has the BiLSTM layer, a dropout layer, a fully connected layer that reduces the dimension of the features, followed by another fully connected layer and the softmax layer.

The classes that are going to be recognized are the written meaning of what is gestured, nonetheless, these annotated classes are not provided.

To deal with this issue, the instances generated within the defined window size (3 frames) are annotated with the corresponding class occupying the data provided in the EAF files associated to each record. Those instances that are not associated to any class

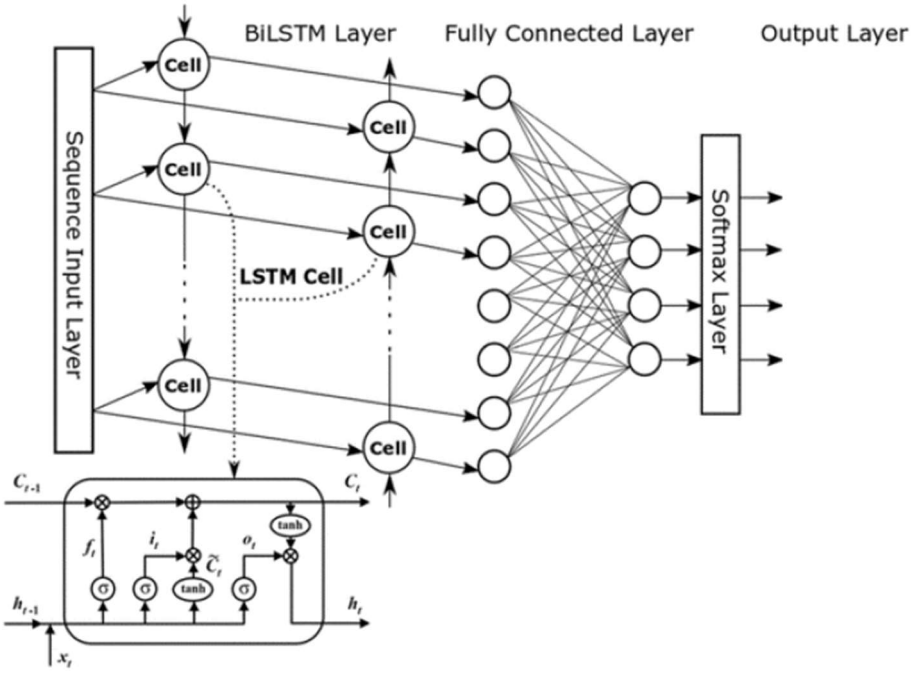


Fig. 5. Base architecture for the recognition task.

are labeled as *blank_transition*, these instances are concerned with transition movements or rest states.

4 Experimental Design and Results

For the experiments carried out, LIBRAS dataset is used. At the current stage of the investigation only 8 videos were used for the tests. In the training process, it was decided to use Google Colab. Available Colab hardware resource is a Tesla T4 graphics processing unit (GPU) that features 16GB RAM, 2,560 NVIDIA CUDA cores, and single-precision performance of 8.1 TFLOPS.

For training and evaluation processes, the data was divided into two sets: 70% for training and 30% for testing. Python language was used in the implementation of the proposal; PyTorch [27] was used for the BiLSTM networks.

The parameters that were defined for all the BiLSTM networks in training process are the number of epochs, which was set to 10 and which was defined empirically through experiments. Also, the batch size is set to 50; the number of cells in each hidden layer is 128; and the learning rate is 0.003.

The experiments were performed by each video using each one of the BiLSTM architectures that were presented in the previous chapter, at the end average accuracy (AvgAcc) and standard deviation (SD) were calculated; Table 1 shows the obtained results. At first glance it can be seen the accuracy is low, to improve the results and

Table 1. Obtained results by each BiLSTM architecture before (BPCA) and after (APCA) PCA process.

Architecture	AvgAcc (BPCA)	SD (BPCA)	AvgAcc (APCA	SD (APCA)
Base	64.79%	9.27%	73.99%	8.24%
Base+ReLU	63.51%	9.73%	72.09%	8.90%
Base+2FC	62.27%	9.46%	71.41%	8.99%

Table 2. Comparison of the proposed method with related work.

Author	Signing data	Accuracy
Amaral et al. [24]	Isolated	88.40%
Passos et al. [23]	Isolated	85.40%
Proposed Work	Continuous	73.99%

analyze the relevance of the features, Principal Components Analysis (PCA) [21] is applied before the recognition process.

PCA was implemented using scikit-learn framework, which gives the option to define the number of components or to define the percentage of the desire variability to preserve in the features. The latter was chosen, this approach used Minka's method [22] to automatically find the number of components, which in this case resulted to be 20. After this, the experiments were performed again and the obtained results showed an improvement, they are shown in the Table 1.

The best result was 73.99%, in order to compare it with related work, it was taken into consideration the review of Wadhawan and Kumar [25], where an extensive analysis by different sign languages of distinct countries was conducted.

This is done because LIBRAS dataset has not been occupied in another sign language recognition works to the best of the knowledge of the authors. Table 2 depicts the comparison of the best obtained result with other author's work, who used distinct datasets employing the same sign language.

Although they have better accuracies for word level recognition, the best result obtained is acceptable and it has the advantage that it was obtained by continuous signing data unlike the works, which have been described as a more complexed and challenging task in section 2.

Interestingly one of the features that it was preserved after the PCA step is the head pose, showing that this feature contributes relevant information. Another relevant finding that it was made is that some classes have between one or three instances and other have more than one hundred, so to improve the results it must be analyzed if data augmentation or imbalance data techniques might help to obtain a more robust recognition model.

Also, as it can be appreciated SD is still high, this needs to be addressed as future work; different difficulty in the sequences or instability in the training process could be the explanation of this behavior. Finally, the difference between the three BiLSTM architectures were so close, this behavior must be analyzed in depth in order to find out why the best result was obtained with the base network.

5 Conclusions and Future Work

In the present work, a proposal for sign language recognition using manual and non-manual features was conveyed. The descriptors were extracted locally to avoid add unnecessary noise in the recognition process, in addition, the relevance of descriptors such as head posture and gaze direction, which have not been used, was analyzed.

The results obtained from the designed experiments are promising, especially if is considered that the dataset used was not acquired and designed for the purpose of sign language recognition. As future work there are several possible actions to be carried out, the first could be to increase the data through data augmentation techniques to obtain a recognition model that has a greater number of instances on those classes that currently have few; in the same topic, the use of imbalance data techniques (subsampling) can also be a path to follow.

Besides that, the implementation of other recognition techniques that are suitable for the problem and that have shown good results, such as Connectionist Temporal Classification (CTC) [29] or Transformers [26], will be carried out. Finally, the design and execution of more experiments considering more data and benchmark datasets will be done to validate the results.

References

1. World health organization: Deafness and hearing loss (2022) www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss
2. Hockett, C. F.: The origin of speech. Scientific American, vol. 203, pp. 88–97 (1960)
3. Elakkiya, R.: Machine learning based sign language recognition: A review and its research frontier. Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 7, pp. 7205–7224 (2020) doi: 10.1007/s12652-020-02396-y
4. Koller, O.: Quantitative survey of the state of the art in sign language recognition (2020) doi: 10.48550/ARXIV.2008.09918
5. Ebrahim-Al-Ahdal, M., Nooritawati, M. T.: Review in sign language recognition systems. In: IEEE Symposium on Computers and Informatics (2012) doi: 10.1109/isci.2012.6222666
6. Fels, S. S., Hinton, G. E.: Glove-talk: a neural network interface between a data-glove and a speech synthesizer. IEEE Transactions on Neural Networks, vol. 4, no. 1, pp. 2–8 (1993) doi: 10.1109/72.182690
7. Grobel, K., Assan, M.: Isolated sign language recognition using hidden Markov models. In: IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, vol. 1, pp. 162–167 (1997) doi: 10.1109/ICSMC.1997.625742
8. Mehdi, S. A., Khan, Y. N.: Sign language recognition using sensor gloves. In: Proceedings of the 9th International Conference on Neural Information Processing, vol. 5, pp. 2204–2206 (2002) doi: 10.1109/ICONIP.2002.1201884
9. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: IEEE International conference on image processing, pp. 1565–1569 (2014) doi: 10.1109/ICIP.2014.7025313
10. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with jointly calibrated leap motion and depth sensor. Multimedia Tools and Applications, vol. 75, pp. 14991–15015 (2016) doi: 10.1007/s11042-015-2451-6

11. Kumar, P., Gauba, H., Roy, P. P., Dogra, D. P.: A multimodal framework for sensor based sign language recognition. Neurocomputing, vol. 259, pp. 21–38 (2017) doi: 10.1016/j.neucom.2016.08.132
12. Ibrahim, N. B., Selim, M. M., Zayed, H. H.: An Automatic Arabic Sign Language Recognition System (ArSLRS). Journal of King Saud University - Computer and Information Sciences, vol. 30, no. 4, pp. 470-477 (2018) doi: 10.1016/j.jksuci.2017.09.007
13. Kong, W. W., Ranganath, S.: Towards subject independent continuous sign language recognition: A segment and merge approach. Pattern Recognition, vol. 47, no. 3, pp. 1294–1308 (2014) doi: 10.1016/j.patcog.2013.09.014
14. Li, K., Zhou, Z., Lee, C. H.: Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications. ACM Transactions on Accessible Computing, vol. 8, no. 2, pp. 1–23 (2016) doi: 10.1145/2850421
15. Elakkiya, R., Selvamani, K.: Extricating manual and non-manual features for subunit level medical sign modelling in automatic sign language classification and recognition. Journal of Medical Systems, vol. 41, no. 11 (2017) doi: 10.1007/s10916-017-0819-z
16. Quadros, R. M., Schmitt, D., Lohn, J., de Arantes Leite, T.: Corpus de libras (2020)
17. Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A.: Ultralytics/yolov5: v4. 0-nn. silu () activations, weights and biases logging, pytorch hub integration. Zenodo (2021)
18. Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., Sheikh, Y.: OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186 (2021) doi: 10.1109/tpami.2019.2929257
19. Baltrusaitis, T., Zadeh, A., Lim, Y. C., Morency, L. P.: OpenFace 2.0: Facial behavior analysis toolkit. In: 13th IEEE International Conference on Automatic Face and Gesture Recognition (2018) doi: 10.1109/fg.2018.00019
20. Aloysius, N., Geetha, M.: Understanding vision-based continuous sign language recognition. Multimedia Tools and Applications, vol. 79, no. 31–32, pp. 22177–22209 (2020) doi: 10.1007/s11042-020-08961-z
21. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, pp. 37–52 (1987) doi: 10.1016/01697439(87)80084-9
22. Minka, T.: Automatic choice of dimensionality for PCA. Advances in Neural Information Processing Systems 13 (2000)
23. Passos, W. L., Araujo, G. M., Gois, J. N., de Lima, A. A.: A gait energy image-based system for brazilian sign language recognition. In: IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 11, pp. 4761–4771 (2021) doi: 10.1109/tcsi.2021.3091001
24. Amaral, L., Ferraz, V., Vieira, T., Vieira, T.: Skelibras: A large 2d skeleton dataset of dynamic brazilian signs. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer International Publishing, pp. 184–193 (2021) doi: 10.1007/978-3-030-93420-0_18
25. Wadhawan, A., Kumar, P.: Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering, vol. 28, no. 3, pp. 785–813 (2019) doi: 10.1007/s11831-019-09384-2
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Conference on Neural Information Processing Systems (2017)
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Proceedings of the 33rd

International Conference on Neural Information Processing Systems, no. 721, pp. 8026-8037 (2019) doi: 10.48550/arXiv.1912.01703
28. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning, vol. 20, no. 3, pp. 273–297 (1995) doi: 10.1007/bf00994018
29. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification. In: Proceedings of the 23rd International Conference on Machine Learning (2006) doi: 10.1145/1143844.1143891

ISSN 1870-4069

Búsqueda armónica binaria para la selección de atributos

Máximo E. Pacheco-Martínez, Maya Carrillo-Ruíz

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

{maximo.pachecomartinez, maya.carrilloruiz}@viep.com.mx

Resumen. La selección de atributos es un proceso crucial en el aprendizaje automático, la cual, ayuda a eliminar información irrelevante, reduciendo el número de atributos y disminuyendo la complejidad del modelo. Debido a que es un problema NP-Completo, se propone usar el algoritmo de búsqueda armónica para llevar a cabo dicha reducción. El proceso descrito en este artículo, consiste en obtener distintos subconjuntos (arreglos binarios) del conjunto original de atributos, evaluar su aptitud y haciendo uso de la búsqueda armónica determinar el mejor de ellos. Este proceso es ejecutado sobre distintas bases de datos de índole médica, utilizando Redes Neuronales Artificiales como clasificador. Los resultados muestran que mediante la modificación al algoritmo de Búsqueda Armónica (Binaria), se logra reducir la dimensionalidad de atributos de las distintas bases médicas. Además, dichos resultados son comparados con otros algoritmos usados para el mismo propósito.

Palabras clave: Búsqueda armónica, búsqueda armónica binaria, redes neuronales artificiales, selección de atributos.

Binary Harmonic Search for Feature Selection

Abstract. Feature selection is a crucial process in machine learning, which helps to eliminate irrelevant information by reducing the number of features and decreasing the model's complexity. Since it is an NP-Complete problem, we propose using the harmonic search algorithm to carry out this reduction. The process described in this article involves obtaining various subsets (binary arrays) from the original feature set, evaluating their fitness, and using harmonic search to determine the best one. This process is executed on various medical databases, using Artificial Neural Networks as a classifier. The results show that by modifying the Harmonic Search (Binary) algorithm, we can reduce the dimensionality of features in different medical databases. Additionally, these results are compared with other algorithms used for the same purpose.

Keywords: Harmonic search, binary harmonic search, artificial neural networks, feature selection.

1. Introducción

La selección de atributos permite reducir la dimensión del problema, obteniendo modelos menos complejos con un porcentaje aceptable en la clasificación. Similar al aprendizaje automático, se divide en dos tipos Supervisado y no Supervisado, únicamente se considera este último para el presente artículo. Los métodos Supervisados, a su vez se dividen en filter, intrinsic y wrapper. Los métodos filter están orientados a la estadística, con el uso del Coeficiente de Pearson y/o Chi cuadrada.

Los métodos de intrinsic a Árboles de Decisión y/o Regularización Lasso. Finalmente, los métodos wrapper a los Algoritmos Genéticos y/o a la Eliminación Recursiva de Atributos. El uso de la búsqueda armónica binaria cae dentro de los métodos wrapper, los cuales consisten en crear modelos a partir de subconjuntos de características.

En la sección 2, se hace un breve análisis de los trabajos relacionados con el tema. En la sección 3, se describe el algoritmo de búsqueda armónica. En la sección 4, se describe la codificación que tendrá la selección de atributos, así como el algoritmo de búsqueda armónica binaria. En la sección 5, se consolidan los conceptos anteriores dando la estructura del modelo que se propone. En la sección 6, se presentan los experimentos. Por último, en la sección 7 se encuentran las conclusiones.

2. Trabajo relacionado

Cuando se tienen problemas donde la función es multimodal, o problemas de combinatoria, donde explorar todas las opciones por fuerza bruta no es viable, se suele recurrir a las metaheurísticas como alternativa.

Algo similar ocurre en el caso de la selección de atributos, donde diversas metaheurísticas son utilizadas para este problema.

Como se mencionó en la primera sección, existen distintos tipos de métodos para la selección de atributos, una variante que engloba varios de ellos es el método GCSA(Genetic Crow Searh Algorithm) [1] que utiliza una combinación de Arboles de Decisión (DT, Decision Tree), Máquina de Soporte Vectorial (SVM, Support Vector Machine), Naive Bayes(NB) y Bosques Aleatorios (RF, Random Forest).

También se ha propuesto para este problema el Algoritmo Binario del Murciélago (BBT, Binary Bat Algorithm), incluso usado para el aprendizaje no supervisado [2]. Otra metaheurística popular es el Enjambre de Partículas (PSO, Particle Swarm Optimization) utilizado para determinar la frecuencia cardiaca fetal [3]. En cuanto a la Búsqueda Armónica, esta se ha aplicado para mejorar el comportamiento de los aparatos auditivos [4].

3. Búsqueda armónica

La búsqueda armónica es una metaheurística inspirada en la armonía musical [5], donde un conjunto de instrumentos emite notas dentro de un rango y tonalidad que al sonar al unísono forman una armonía. Similar a todas las metaheurísticas, se parte de una población inicial o en este caso llamada Memoria Armónica (HM, Harmony

Tabla 1. Representación de un subconjunto de atributos.

F_1	F_2	F_3	F_4	F_5	F_6
0	1	1	0	0	1

Memory), donde cada vector o individuo de la población representa las notas que toca un instrumento. El tamaño de la memoria armónica se denota como HMS (Harmony Memory Size) y la dimensión del problema como m:

$$HM = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{HMS1} & x_{HMS2} & \cdots & x_{HMSm} \end{bmatrix}. \quad (1)$$

Para producir una nueva armonía (un individuo $\mathbf{x}^{new}$), se toma un elemento al azar de la memoria armónica, a este proceso se le denomina **consideración de la memoria armónica** (HMC, Harmony Memory Consideration). A grandes rasgos, es, generar un número aleatorio (random) y compararlo respecto a un umbral de consideración (HMCR Harmony Memory Consideration Rate), si es menor, entonces el elemento j del individuo i será dicho valor de la memoria armónica. Es decir:

$$\text{si random} < \text{HMCR: } x_{ij}^{\text{new}} = x_{ij}, \quad x_{ij} \in \{x_{1j}, x_{2j}, x_{3j}, \cdots, x_{\text{HMS}j}\}. \quad (2)$$

Si no se cumple, se hace una **inicialización aleatoria** del elemento dentro de un rango específico dado por el límite inferior l y superior u del problema:

$$x_{ij}^{\text{new}} = l + \text{random} \cdot (u - l). \quad (3)$$

Una vez obtenido el elemento, se realiza una **corrección de tono** (PA, Pitch Adjusting), esto, nuevamente mediante un aleatorio y una comparativa respecto a un umbral de corrección de tono (PAR, Pitch Adjustment Rate). Para controlar el ajuste, se usa un ancho de banda (BW, Bandwidth). Es decir:

$$\text{si random} < \text{PAR: } x_{ij}^{\text{new}} = x_{ij} \pm \text{random} \cdot \text{BW}. \quad (4)$$

Una vez calculado todos los elementos de $\mathbf{x}^{new}$, el siguiente paso es **actualizar la memoria armónica**, para ello, es necesario evaluar la aptitud del nuevo individuo y verificar si es mejor que el peor existente, de ser así, el anterior es reemplazado por el nuevo. El Algoritmo 1 ilustra todos los pasos necesarios.

4. Representación para la selección de atributos y búsqueda armónica binaria

En este trabajo se utiliza la búsqueda armónica para la selección de atributos, para ello, es necesario definir primero la codificación que tendrán los arreglos de la memoria armónica. Una vez dada esta codificación, la búsqueda armónica tal cual el Algoritmo 1 necesitará ser modificada.

4.1 Codificación

Considere un problema donde se cuenta con 6 atributos, cada uno de estos será marcado con un 1 cuando sea un atributo a ocupar, o 0 en caso contrario. Es decir, se tiene una

codificación binaria. Continuando con el ejemplo, una posible selección de atributos dentro de 6 atributos posibles es la siguiente:
donde se observa que, del conjunto de atributos original, solo los atributos F_2, F_3 y F_6 serán ocupados.

4.2 Búsqueda armónica binaria

El Algoritmo 1, está pensado para problemas con números reales, por ello, es necesario realizarle modificaciones para adaptarlo a la codificación binaria. La variación más sencilla [6], consiste en, remover el paso de ajuste de tono, para quitarse la introducción del aleatorio real y el ancho de banda, y cambiar la inicialización aleatoria a una binaria, pues no es necesario obtener un punto entre los limites superior e inferior y solo se necesita seleccionar aleatoriamente un 1 o un 0. La operación de selección de aleatorio binario se denota como *random*(0|1). La búsqueda armónica con estos ajustes se muestra en el Algoritmo 2.

Algoritmo 1. Búsqueda Armónica

```
Require: HMS, HMCR, PAR, BW, NI (no. iteraciones)
Ensure: Todos los aleatorios dentro de (0,1)
  for i = 1 : i <= HMS do
    Genera el vector aleatorio x_i de la memoria.
    Evalúa la aptitud del individuo f(x_i).
  end for
  for t = 1 : t <= NI do
    for i = 1: i <= HMS do
      for j = 1 : j <= m do
        if random < HMCR then          ▷ Memory Consideration
          x_ij^new = x_ij,  x_ij ∈ {x_1j, x_2j, x_3j, ⋯, x_HMSj}  ▷ i es aleatorio
        if random < PAR then           ▷ Pitch Adjustment
        x_ij^new = x_ij ± random · BW
          end if
          if x_ij^new < l then         ▷ Evita estar fuera de rango
            x_ij^new = l
          end if
          if x_ij^new > u then         ▷ Evita estar fuera de rango
            x_ij^new = u
          end if
        else
          x_ij^new = l + random · (u − l)    ▷ Inicialización Aleatoria
        end if
      end for
    end for
    if f(x^new) < f(x^worst) then      ▷ Actualización de la Memoria
      x^worst = x^new
    end if
  end for
```

5. Estructura del modelo

En la sección anterior, se introdujo la codificación, así como la modificación a la búsqueda armónica para poder trabajar con arreglos binarios, con esto, es posible utilizar el Algoritmo 2 tal cual se ilustra, sin embargo, hay un componente pendiente por definir, y este es la función de aptitud.

Esta función depende del problema que se quiera resolver, en el presente trabajo, se hace uso de bases de datos de índole médica, las cuales tienen como objetivo clasificar ciertos padecimientos, es por ello, que la función de aptitud debe ser un clasificador. El utilizado en este artículo es una Red Neuronal, la cual genera un clasificador por cada subconjunto de atributos contenidos en la memoria armónica.

Cada uno de los clasificadores, es entrenado solo con los atributos seleccionados, dando como valor de aptitud el error calculado mediante la entropía cruzada. La red neuronal, tiene un número de neuronas en la capa de entrada igual al número de atributos seleccionados, mientras las neuronas en la capa de salida dependen del problema, es decir, del número de clases posibles de cada base.

El ejemplo de una estructura, se puede apreciar en la Figura 1, el problema en cuestión cuenta con dos clases y en cada clasificador se mantiene constante, mientras el número de neuronas de entrada varía entre uno y seis dependiendo de los atributos seleccionados en color negro. Finalmente, cada clasificador arroja cierto error, que se intenta minimizar, considerado como la aptitud (costo).

Algoritmo 2. Búsqueda Armónica Binaria

```
Require: HMS, HMCR, PAR, BW, NI (no. iteraciones)
Ensure: Todos los aleatorios dentro de (0,1)
  for i = 1 : i <= HMS do
    Genera el vector aleatorio x_i de la memoria.
    Evalúa la aptitud del individuo f(x_i).
  end for
  for t = 1 : t   <= NI do
    for i = 1 : i <= HMS do
      for j = 1 : j <=   m do
        if random < HMCR then        ▷ Memory Consideration
          x_ij^new = x_ij,  x_ij ∈ {x_1j, x_2j, x_3j, ⋯, x_HMSj} ▷ i es aleatorio
        else
          x_ij^new = random(0|1)     ▷ Inicialización Aleatoria
        end if
      end for
    end for
    if f(x^new) < f(x^worst) then    ▷ Actualización de la Memoria
      x^worst = x^new
    end if
  end for
```

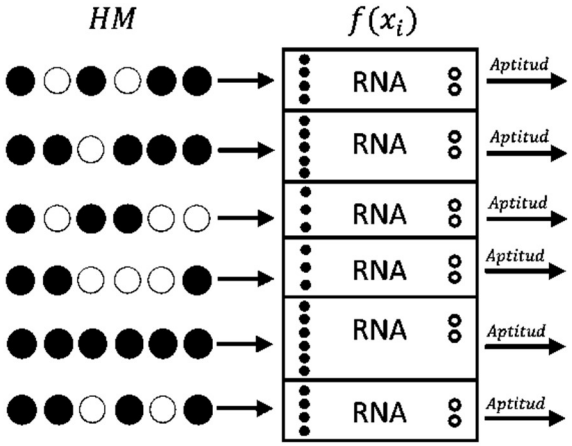


Fig. 1. Estructura del modelo para un problema de dos clases.

Tabla 2. Información de las bases de datos.

Base	Atributos	Instancias	Clases
Lung Cancer [7]	57	32	3
Hepatitis [8]	20	155	2
Dermatology [9]	35	366	6
Pima Indians Diabetes [10]	9	768	2

6. Experimentos

El lenguaje utilizado para la implementación, fue Julia, el cual, es un lenguaje compilado que permite tener velocidades de ejecución mayores a los lenguajes interpretados. El algoritmo de búsqueda binaria fue programado desde cero, mientras para la red neuronal, se usó la biblioteca FLUX, la cual nos permite crear redes neuronales de manera robusta y sencilla.

El algoritmo generado se denota como HS+ANN. El modelo se probó sobre cinco bases de datos distintas, donde el número de atributos, instancias (número de registros) y clases se da en la Tabla 2. Dentro de los atributos se incluye la clase a la que pertenece cada instancia. De cada base se tomó el 80% para el conjunto de entrenamiento y 20% para el conjunto de prueba.

Para el procesamiento de datos, todas las bases fueron normalizadas y los datos faltantes en alguna instancia fueron reemplazados por ceros, además, en todos los experimentos se incluye el vector de unos, es decir, el vector donde se ocupan todos los atributos excluyendo la clase. La Red Neuronal consiste de dos capas ocultas, con un número de neuronas por capa definido como:

$$\text{neuronas ocultas} = 2 \cdot N + 1, \tag{5}$$

Tabla 3. Resultados de los 100 experimentos por base de datos.

Variable	Media	Min	Med	Max
exentre	1	1	1	1
exprueb	0.5	0.22	0.55	0.66
noatrib	34	27	34	56

(a) Lung Cancer

Variable	Media	Min	Med	Max
exentre	1	1	1	1
exprueb	0.78	0.67	0.77	0.87
noatrib	15	10	15	19

(b) Hepatitis

Variable	Media	Min	Med	Max
exentre	1	1	1	1
exprueb	0.963	0.92	0.96	0.986
noatrib	28	22	29	33

(c) Dermatology

Variable	Media	Min	Med	Max
exentre	0.788	0.76	0.788	0.827
exprueb	0.713	0.649	0.72	0.766
noatrib	7.7	6	8	8

(d) Pima Indians Diabetes

donde N es el número de características del modelo a ser probado. Con función de activación relu en las capas ocultas y softmax en la capa de salida. Finalmente, la función de costo a minimizar es la entropía cruzada. El número de épocas se fijó en 150 y la tasa de aprendizaje en 0.01, ambos parámetros se mantienen constantes en todas las pruebas y bases de datos.

Con respecto a los parámetros de la búsqueda armónica, estos son los siguientes: HMS = 20, HMCR = .95, NI = 100 y una tolerancia de iteraciones sin mejora del 10%. Similar a los parámetros anteriores, se mantienen constantes en todas las pruebas y bases de datos. En cada experimento se determina el mejor clasificador de acuerdo con la entropía cruzada y el conjunto de atributos que lo generó se guarda como el mejor, así como la exactitud de entrenamiento y pruebas. Dicho experimento se repite un total de 100 veces para obtener los 100 mejores conjuntos de atributos. Para elegir el mejor de estos 100, se toma aquel que pondera la mejor exactitud tanto en el entrenamiento como en las pruebas usando la regla:

$$\text{exactitud} = \text{exactitud de entrenamiento} \cdot \text{exactitud de pruebas}. \quad (6)$$

Los resultados estadísticos de las 100 ejecuciones se muestran en la Tabla 3, donde se da: la media, mediana (med), mínimo (min) y máximo (max) para la exactitud del entrenamiento (exentre), exactitud de las pruebas (exprueb) y el número de atributos (noatrib). La información del mejor obtenido dentro de estos 100 se encuentra en Tabla 4, donde, los atributos originales corresponden al número total de atributos menos el atributo de la clase a la que pertenece cada instancia.

Atributos seleccionados, es el número de atributos utilizados por el mejor experimento y por ende el resultado final. Atributos reducidos es solo la cantidad en la que se redujo la dimensión del problema (originales - seleccionados). Los atributos elegidos para cada uno de los mejores, se encuentran en la Tabla 5.

Tanto la base Lung Cancer, Hepatitis y Dermatology, en los 100 experimentos, la exactitud en el entrenamiento es del 100%, esto se puede afirmar pues no es posible tener valores en la exactitud por arriba de 1, así, una media 1 efectivamente representa

Table 4. Información del mejor experimento para cada base de datos empleando exactitud

Variable	Lung Cancer	Hepatitis	Dermatology	Pima I. Diabetes
exentre	1	1	1	0.788
exprueb	0.66	0.87	0.986	0.766
atrib. originales (sin clase)	56	19	34	8
atrib. seleccionados	34	16	28	7
atrib. reducidos	21	3	6	1

Table 5. Características elegidas para el mejor experimento por base de datos.

Base	Características
Lung Cancer	$P1$, $P2$, $P3$, $P5$, $P6$, $P7$, $P8$, $P13$, $P14$, $P16$, $P17$, $P18$, $P19$, $P20$, $P21$, $P23$, $P24$, $P27$, $P28$, $P29$, $P30$, $P34$, $P38$, $P39$, $P40$, $P42$, $P43$, $P44$, $P46$, $P48$, $P50$, $P51$, $P54$, $P56$
Hepatitis	Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Liverfirm, Spleenpalpable, Spiders, Ascites, Varices, Alkphosphate, Sgot, Albumin, Protime, Histology
Dermatology	erythema, definite_borders, itching, koebner_phenomenon, polygonal_papules, follicular_papules, oral_mucosal_involvement, knee_and_elbow_involvement, scalp_involvement, family_history, eosinophils_in_the_infiltrate, pnl_infiltrate, fibrosis_of_the_papillary_dermis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing_of_the_rete_ridges, elongation_of_the_rete_ridges, thinning_of_the_suprapapillary_epidermis, focal_hypergranulosis, disappearance_of_the_granular_layer, vacuolisation_and_damage_of_basal_layer, spongiosis, saw-tooth_appearance_of_retes, follicular_horn_plug, perifollicular_parakeratosis, band-like_infiltrate
Pima Indians Diabetes	Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Age

el 100% en todos los experimentos. El número medio de atributos por base es de 34 para la primera, 15 para la segunda y 28 para la tercera, los cuales son un número de atributos similares al mejor experimento de cada base dado en la Tabla 4, donde el único con un valor diferente es el de Hepatitis.

Sin embargo, esto no ocurre para la exactitud en las pruebas para Lung Cancer y Hepatitis, en esta ocasión, la media es de 0.5 cuando el mejor tiene una exactitud de 0.66 y media 0.78 con valor del mejor 0.87 respectivamente, para el caso de Dermatology sí se cumple, donde 0.963 es cercano a 0.986. En todos los casos, el mejor experimento está por arriba de la media, esto refuerza lo bien conocido, que para este tipo de heurísticas el número de experimentos es importante, puesto que, cada ejecución depende de los aleatorios que genere la máquina, dando resultados distintos.

La base **Pima Indians Diabetes,** se mantiene coherente con sus medias y los resultados del mejor, sin embargo, es la clase que menos reducción de atributos tiene,

Table 6. Comparativa directa con [1] empleando exactitud.

Base	Caract.	Instancias	Clases	Atributos Seleccionados	exprueb
Lung Cancer	56/55	32/31	3/2	34/26	66/81.30
Hepatitis	19/18	155/153	2/2	16/9	87/80.01
Dermatology	34/32	366/364	6/6	28/23	98.6/97.54
Pima Indians Diabetes	8/8	768/765	2/2	7/5	76.6/89.40

Table 7. Comparativa directa con otros algoritmos incluidos en [1] empleando exactitud.

Base	Bagging with C4.5 method	Boosting with C4.5 method	ACO bagging	ACO boosting	GCSA with DCNN	HS with ANN
Lung Cancer	84.20	85.45	77.34	77.34	**88.23**	66.00
Hepatitis	74.50	74.28	79.30	77.80	80.09	**87.00**
Dermatology	96.20	96.50	97.55	97.55	97.12	**98.60**
Pima I. Diabetes	78.34	75.10	**89.45**	87.45	89.34	76.60

solamente disminuyendo en 1, si bien, Hepatitis también reduce poco comparado con su número total de variables, en su media representa un valor de 15, mientras la media para Pima Indians Diabetes es cercana a 8, que es prácticamente sin cambios, siendo solo el mejor el que logra la reducción de 1. En cualquiera de los casos, la búsqueda armónica permite la reducción de atributos en todas las bases presentadas.

En la Tabla 6, se muestra una comparativa con el trabajo previo directo dado en [1], en el cual, se utilizan las mismas bases. Cabe resaltar, que, a pesar de ser las mismas, existen algunas diferencias, muy posiblemente dadas por la época entre el presente artículo y el artículo de [1], además, en [1] no se específica el "%" de datos ocupados para el entrenamiento y pruebas, por lo que una comparativa estricta no es posible, sin embargo, sirve como referencia para ubicar el presente trabajo con los realizados previamente.

La información se presenta separada por una "/", donde, al lado izquierdo se ubica la información de este artículo, y a la derecha, la información en [1]. En todos los casos, el algoritmo presentado en [1] (GCSA+DCNN), resulta en una reducción mayor de atributos, sin embargo, en la exactitud, únicamente rebasa al algoritmo presentado en este artículo en dos experimentos, por lo tanto, una reducción mayor de atributos no garantiza una mayor exactitud. Finalmente, un comparativo de exactitud obtenida con diferentes algoritmos, también ofrecida por [1], se muestra en la tabla 7, donde se puede ubicar al HS+ANN junto a otros cinco algoritmos.

7. Conclusiones

Las metaheurísticas son de utilidad para la selección de atributos, dentro de esta gama de algoritmos, la búsqueda armónica puede ser utilizada mediante la variación

binaria. Los resultados fueron presentados en la sección 6, donde se aprecia la reducción obtenida, esto, siempre y cuando se conjunte con un buen algoritmo de clasificación, como lo son las Redes Neuronales, con las que, trabajando en conjunto, se logran buenos resultados en la exactitud del entrenamiento, variando un poco en el de pruebas.

Este último problema está ligado a los datos de cada base, por lo que un análisis mayor a cada una o variación del clasificador será necesario para obtener mejores resultados. Como trabajo futuro, se podría realizar un análisis a profundidad sobre la presencia de ciertos atributos dentro de los 100 experimentos realizados, para tener un panorama mejor de los atributos de mayor relevancia y los cuales deben de estar presentes en cualquier modelo.

Además, se puede trabajar en cada base por separado para lograr ajustar de mejor manera los parámetros de la red neuronal y evitar un sobreajuste, recordando que en todas las bases se mantuvo la misma estructura y parámetros. Cabe destacar, que como cualquier heurística requirió de un número alto de iteraciones para obtener los resultados mostrados. Pese a que las pruebas contenían el vector de unos, los modelos generados resultan con atributos reducidos logrando el objetivo principal de reducir atributos empleando búsqueda armónica.

Referencias

1. Nagarajan, S. M., Muthukumaran, V., Murugesan, R., Joseph, R. B., Meram, M., Prathik, A.: Innovative feature selection and classification model for heart disease prediction. Journal of Reliable Intelligent Environments, vol. 8, no. 4, pp. 333–343 (2021) doi: 10.1007/s40860-021-00152-3
2. Rani, A. S. S., Rajalaxmi, R. R.: Unsupervised feature selection using binary bat algorithm. In: 2nd International Conference on Electronics and Communication Systems, pp. 451–456 (2015) doi: 10.1109/ecs.2015.7124945
3. Inbarani, H. H., Banu, P. K. N., Azar, A. T.: Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Computing and Applications, vol. 25, no. 3–4, pp. 793–806 (2014) doi: 10.1007/s00521-014-1552-x
4. Geem, Z. W.: Music-inspired harmony search algorithm: theory and applications. vol. 191, Springer (2009)
5. Geem, Z. W., Kim, J. H., Loganathan, G.: A new heuristic optimization algorithm: Harmony search. Simulation, vol. 76, no. 2, pp. 60–68 (2001) doi: 10.1177/003754970107600201
6. Kong, X., Gao, L., Ouyang, H., Li, S.: A simplified binary harmony search algorithm for large scale 0-1 knapsack problems. Expert systems with applications, vol. 42, no. 12, pp. 5337–5355 (2015) doi: 10.1016/j.eswa.2015.02.015
7. Hong, Z. Q., Yang, J. Y.: Lung Cancer. UCI Machine Learning Repository (1992) archive.ics.uci.edu/ml/datasets/lung+cancer
8. Hepatitis. UCI Machine Learning Repository (1988) archive.ics.uci.edu/ml/data sets/hepatitis
9. Dermatology. Homepage (2022) datahub.io/machine-learning/dermatology
10. Pima Indians diabetes database. UCI Machine Learning (2016) www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download

ISSN 1870-4069

Entrenamiento de un clasificador de videos DeepFake en un equipo de cómputo con recursos limitados

Odón D. Carrasco-Limón, Maya Carrillo-Ruiz,
María de Lourdes Sandoval-Solis

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

{odond.carrasco, maya.carrilloruiz,
mariad.sandovalsolis}@viep.com.mx

Resumen. En la actualidad la utilización de aplicaciones móviles para generar videos DeepFake al alcance de cualquiera, suscita la problemática de no poseer herramientas que nos permitan discernir si un video es o no un DeepFake, por lo que surge la necesidad de poner al alcance de usuarios técnicas para su detección. En este documento se presenta un método, para la detección de videos modificados por la técnica DeepFake. Dicho método utiliza redes neuronales que pueden ser entrenadas en ordenadores personales, por medio de una técnica que denominamos aprendizaje particionado. Los resultados obtenidos muestran que la exactitud obtenida con el aprendizaje particionado es aceptable además de reducirse considerablemente el tiempo de entrenamiento.

Palabras clave: Visión por computadora, clasificadores, CNN, aprendizaje particionado, deepfake.

Training a DeepFake Video Classifier on a Computer with Limited Resources

Abstract. Nowadays, the use of mobile applications to generate DeepFake videos accessible to anyone raises the issue of not having tools that allow us to discern whether a video is a DeepFake or not. Therefore, there is a need to provide users with techniques for its detection. This document presents a method for detecting videos modified by the DeepFake technique. This method utilizes neural networks that can be trained on personal computers using a technique called partitioned learning. The results obtained show that the accuracy achieved with partitioned learning is acceptable, and the training time is significantly reduced.

Keywords: Computer vision, classifiers, CNN, partitioned learning, deepfake.

1. Introducción

Actualmente existen aplicaciones para dispositivos móviles que generan videos DeepFake [1], técnica de intercambio de identidad, de manera completamente automática. Sin embargo, no existen herramientas que nos permitan detectar dicha manipulación.

Aunque se encuentra diversas aproximaciones como las generadas en el concurso de detección de DeepFake (DFDC) [2], estas soluciones requieren de poder de cómputo considerable para el entrenamiento de redes neuronales convolucionales (CNN).

Dado que no siempre se cuenta con tal poder de cómputo, el presente trabajo propone un método de detección de videos falsos que puede ser entrenado en computadoras personales para evitar posibles estafas y manipulación de la información.

Desde el 2017 que apareció este término, se encontraron más de 14,698 videos manipulados, como menciona la cadena de noticias BBC en su reportaje del 2019[3]. Además, del total de videos DeepFake encontrados en la red en ese momento, el 96% eran de naturaleza pornográfica. Ante estos eventos, métodos como el propuesto, día a día serán de mayor utilidad.

En la sección 2, se presenta tres aproximaciones del estado del arte, se describen brevemente sus métodos y a partir de ellas se esquematiza un modelo general. En la sección 3 se describen las etapas del método propuesto, así como las herramientas utilizadas. En la sección 4, se describe el conjunto de datos empleado, el procesamiento realizado y su organización, para emplearlo en aprendizaje automático. En la sección 5 se analizan los resultados de los experimentos y finalmente en la sección 6 se presentan las conclusiones y trabajo futuro.

2. Trabajos relacionados

Tolosana et al. en [1] describen 4 técnicas de manipulación facial en videos, las cuales son: a) Síntesis facial completa, la cual crea rostros inexistentes realistas por medio de poderosas redes adversarias generativas (GAN). b) Intercambio de identidad, la cual cambia el rostro de una persona por otro, mediante técnicas de aprendizaje profundo también conocido por DeepFake. c) Manipulación de atributos, que cambia características como: color de pelo, color de ojos, edad, añade gafas, etc., por medio de redes GAN. d) Intercambio de expresiones, esta manipulación cambia las expresiones de una persona por la de otra, esta es la técnica más reciente con técnicas de Neural-Textures, mediante poderosas redes GAN. De estas técnicas la más accesible al público es la de DeepFake existiendo aplicaciones móviles que permiten crear videos modificados de manera automática, por esto se buscará un detector para esta técnica.

A continuación, se describen brevemente los trabajos más relevantes, en los que podrá observarse que existen etapas comunes en las soluciones propuestas, las cuales son: una etapa de extracción de rostros en la cual se separa el rostro para determinar si un video es real o no, una etapa de extracción de características en la cual típicamente se utiliza alguna red CNN para que las determine y una etapa de clasificación donde a partir de los resultados de la extracción de características se determina si el video es real o no, dichas etapas se pueden observar en la Figura 1.

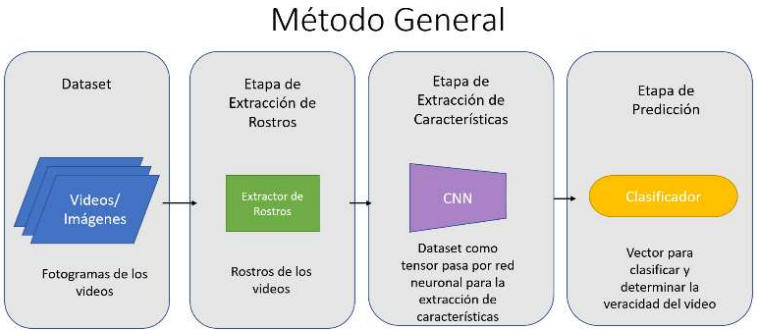


Fig. 1. Etapas del modelo general para detección de DeepFake. Creación propia.

Tolosana et al. en [4], parten de otros estudios que consideran características de bajo y alto nivel en el rostro del individuo, establecen que la región de los ojos, la boca y la pose de la cabeza, dan muchas características para discernir si el video es falso o no, así decide extraer estas 4 regiones y en la última coloca el resto del rostro.

Toma todos los fotogramas del video, y obtiene los rostros mediante el toolkit OpenFace2 [5], continúa con una etapa de extracción de características utilizando dos Redes CNN, una XceptionNet preentrenada con pesos de ImageNet [4] y una red cápsula VGG19. A continuación mediante reglas de inferencia obtiene el área bajo la curva (AUC) de la gráfica Receiver Operating Characteristic (ROC), de 0.994 para la base de datos FF ++, 0.91 para la base de datos DFDC y 0.836 para la base de datos de Celef-db.

El equipo ganador del concurso DFDC, fue Seferbekov, con una propuesta que aparece en GitHub como Selimsef [6] del año 2020, la cual consta de una etapa de extracción de todos los fotogramas del video original, extracción de rostros por medio de una red CNN, aplicación de aumentos a estos rostros, como ruido gaussiano, transformaciones mediante rotación, escalamiento, escala de grises etc.

Para la etapa de extracción de características, el autor propone EfficientNet [7], que mejora al preentrenarla con Noisy Student (una aproximación de aprendizaje semi-supervisados), finalmente realiza la clasificación por medio de reglas de inferencia, logrando un desempeño con Log Loss[1] de 0.42798.

El autor ganador del segundo lugar en DFDC, H. Zhao y su equipo WM [8], en la etapa de extracción de rostros utilizaron una red CNN RetinaFace, aplican aumentos a los rostros extraídos, posteriormente en la etapa de extracción de características utilizan una XceptionNet preentrenada, EfficientNet y XceptionNet con una red de aumento de datos (WSDAN). Crean una interpolación bilineal con estas tres redes, obteniendo un desempeño con Log Loss de 0.42842.

Li et al. en [11] utilizan Dlib como extractor de rostros, en la etapa de extracción de características utilizan una Red Resnet101, finalmente en la clasificación utilizan reglas de inferencia midiendo el desempeño en AUC en el conjunto UADFV obtienen 97.4.

Si bien los trabajos que se presentan no muestran el tiempo que tomó el entrenamiento de cada modelo, cada uno intenta tener la mejor exactitud ocupando los

[1] Métrica específica de desempeño, la fórmula de Log Loss es -1 * el logaritmo de la función de probabilidad, para cualquier problema dado, un valor de pérdida logarítmica más bajo significa mejores predicciones

Tabla 1. Tabla comparativa del estado del arte.

Articulo	Extractor de Rostros	Extractor de Características	DB	Medida de Desempeño
Tolosana- Facial Regions Features (2020)[4]	OpenFace2+ Alineamiento	Xception P. y Capsule Network (VGG19)	FF ++ | DFDC | Celb-DF	0.994 | 0.91 |0.836
S. Seferbekov - Selimsef-(2020)[5]	MTCNN + Aumentos	EfficientNet B7 P. con Noisy Student	DFDC	0.42798 LogLoss
H. Zhao -WM- (2020)[7]	RetinaFace + Alineamiento + Aumentos	Xception P. y EfficientNet B3 + WSDAN y Xception P. + WSDAN	DFDC	0.42842 LogLoss
Li-Face Warping Features (2019)[11]	Dlib	ResNet101	UADFV	0.974

mejores algoritmos y herramientas disponibles en su momento, mismas que implícitamente buscan una mejora en tiempo.

3. Arquitectura del sistema

Para el análisis a profundidad de las etapas mencionadas en la sección anterior, se compararon los métodos utilizados por los autores como se muestra en la Tabla 1. Así se identificaron el extractor de rostros, el extractor de características y el conjunto de datos adecuado para el entrenamiento.Para la selección se eligieron las herramientas con mayor exactitud. A continuación, se describe brevemente en qué consisten las herramientas seleccionadas de la Tabla 1.

Openface2, es un extractor de rostros seleccionado para este trabajo, se compone de una serie de herramientas las cuales implementan los mejores algoritmos para la estimación de Landmarks, utilizando una CE-CLM (CNN de expertos en cascada), incluso en imágenes con rostros de perfil y zonas no visibles.

Openface2 tiene algunos errores en videos, por lo que es recomendable utilizar el extractor dos veces ocupando la salida del primero como la entrada al segundo, para reducir los errores. A pesar de la existencia de estos fallos, Openface2 logra ser el mejor extractor de rostros según [4] en el cual se comparan diversos extractores de rostros.

3.1 Extracción de características

En el campo de visión por computadora se han creado diversas arquitecturas de CNN y se han puesto a prueba en el conjunto de datos ImageNet con el fin de detectar objetos en imágenes. Tales CNN han sido ResNet, InceptionNet y XceptionNet, por dar algunos ejemplos.

Todas estas redes buscan tener mayor precisión en la detección de objetos, trabajando con redes robustas incrementando continuamente su tamaño. Al escalar la red en profundidad, ancho (tamaño de vector de salida) o tamaño de canal (tamaño de vector de entrada); se obtiene mayor precisión. Sin embargo, el tiempo de entrenamiento se incrementa, así como el número de parámetros y el coste

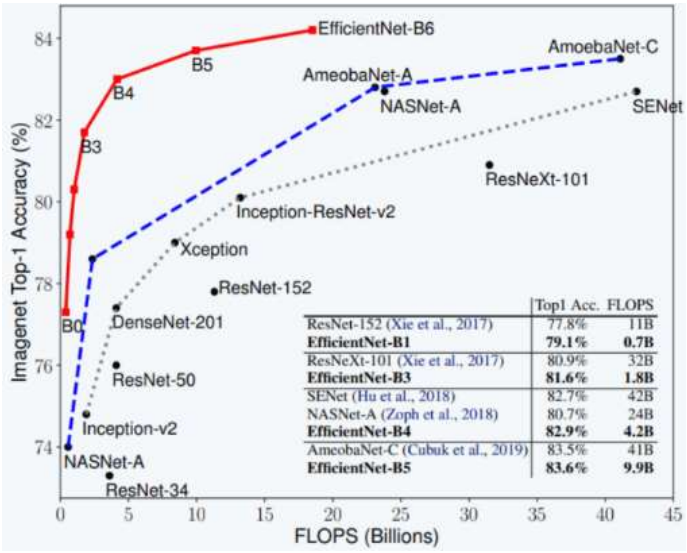


Fig. 2. Comparación de la Familia de CNN EfficientNet comparada con otras CNN en número de operaciones de punto flotante, extraída de [7].

computacional. La mayoría de las CNN buscan escalar un solo valor de los mencionados, pero nunca una combinación de estos [7].

Por otro lado, EfficientNet surge como una red que intenta encontrar un factor de escalamiento para los parámetros antes mencionados, buscando maximizar la precisión, minimizando el número de parámetros y el coste computacional. Esta red se modeló en forma de función, y de manera empírica se encontraron valores para escalar la red, logrando una exactitud superior a las CNN mencionadas previamente.

Con estos factores de escala, los autores propusieron un caso base y aplicando estos factores en forma recursiva crearon una familia de CNN llamada EfficientNet. B0 es el caso base, B1 el siguiente factor, y así sucesivamente hasta el B7, teniendo similar o mejor exactitud que otras CNN con un número menor de parámetros y número de operaciones de punto flotante como se puede ver en la Figura 2.

En el concurso DFDC los ganadores utilizan EfficientNet como extractor de características demostrando que esta familia de arquitecturas no solo logra discernir objetos en imágenes, si no también logra extraer características en rostros humanos.

Esta disminución de parámetros y la reducción significativa de operación de punto flotante hacen que EfficientNet, como extractor de características, pueda utilizarse en cualquier equipo. Aunque la disminución de poder de cómputo requerido no es suficiente para utilizar los modelos más grandes en equipos personales.

3.2 Problemas por tamaño de información en extracción de características

Según lo observado en la Figura 2 utilizando una arquitectura de familia más grande se obtendrá más exactitud. El problema es que el tamaño necesario para procesar la información crece, por ejemplo, en B0 la entrada requiere que las imágenes tengan un tamaño de 224 pixeles de largo y ancho. Como la imagen tiene 3 canales (RGB), el

Tabla 2. Crecimiento de potencia para EfficientNet.

Base modelo	Resolución	Potencia necesaria para 34,778 imágenes
EF B0	224	4.88GB +modelo
EF B1	240	5.60 GB +modelo
EF B2	260	6.57 GB +modelo
EF B3	300	8.75 GB +modelo
EF B4	380	14.03 GB +modelo
EF B5	456	20.20 GB +modelo
EF B6	528	27.09 GB +modelo
EF B7	600	34.98 GB +modelo

tamaño del tensor para entrenar la red tendría un tamaño de 224* 224 * 3 * número de imágenes en la base de datos ocupadas para entrenar.

Así el poder de cómputo necesario para poder manejar el tensor crece rápidamente al tener una base de datos con más imágenes y al cambiar la arquitectura como se puede observar en la Tabla 2, lo que vuelve inviable el manejo del tensor entero para computadoras personales.

Suponiendo que la base de datos tenga una cantidad de 34,778 imágenes, como es la base de UADFV tras extraer los rostros, entonces se necesitará la potencia de cómputo que se observa en la Tabla 2.

Dado que pocas computadoras personales cuentan con tal potencia de cómputo para trabajar con el tensor entero, existen alternativas para poder trabajar sin contar con dicha potencia. Una alternativa es incrementar la potencia del equipo con el cual se entrena la red, pero muchas veces esto no es posible. Por lo que se propone el Aprendizaje particionado.

3.3 Aprendizaje particionado

En el presente trabajo como alternativa para operar con tensores (conjuntos de datos) que superen la potencia de cómputo disponible se propone ocupar lo que llamamos Aprendizaje particionado (AP). El AP se realiza dividiendo un tensor para la resolución del problema, entrenando la red con segmentos del tensor y ocupando los modelos resultantes con otros segmentos distintos del mismo, para reducir la potencia de cómputo necesaria y así disminuir el tiempo de entrenamiento. En la Tabla 3 se mencionan el algoritmo de AP.

4. Experimentos

El objetivo será medir la exactitud que alcanza una red con la técnica AP comparada con una red que no utiliza AP, dado que el conjunto de datos esta balanceado.

Las características del equipo con el que se realizaron las pruebas son las siguientes, la computadora tiene un procesador Ryzen 5 2600, cuenta con 32 Gb de memoria RAM, tiene una tarjeta de video dedicada Nvidea RTX 3070 con 8Gb de VRAM y posee una unidad de almacenamiento de 240Gb de espacio, el lenguaje utilizado para programar fue Python.

Tabla 3. Algoritmo de AP.

1	Partir el tensor T en dos subconjuntos A y B, asignándoles el D% e ID% de datos respectivamente.
2	Crear un modelo M entrenado en 15 épocas, con el D% de datos del subconjunto A.
3	Utilizar el ID% de datos del subconjunto B para entrenar el modelo M, en 5 épocas. Y producir los Modelos D% + ID%

Tabla 4. Nombre de los modelos y cantidad de imágenes.

Nombre Modelo	Imágenes
EF0 10%	3477
EF0 20%	6956
EF0 50%	17389
EF0 100%	34778

Aunque las características del equipo ocupado parecieran alejadas a lo que se pudiera entender por computadoras personales, los requerimientos de hardware que mencionan los equipos de Seferbekov[6] y Zhao H. et. al [8] hacen inviable el entrenamiento de sus modelos con una sola tarjeta de video dedicada, misma razón por la cual surge la idea de ocupar el AP. A continuación, se describe el dataset utilizado para las pruebas.

4.1 Dataset

Para la técnica de intercambio de identidad existen dos generaciones de conjuntos prueba: la primera generación, con pocos videos de baja resolución y con personas hablando a la cámara directamente; la segunda generación, con videos de resoluciones más altas, diversos contextos de iluminación, profundidad y poses de personas.

Con la finalidad de iniciar la experimentación se seleccionó base de datos UADFV, que es comparable con la base de datos FF++ por ser de primera generación, y obtener un AUC de 99.4% en ambas.

UADFV es una colección de 98 videos etiquetados, 49 reales y 49 modificados por la técnica de intercambio de identidad DeepFake, creada por Y. Li et al. [10], cada video tiene una duración de entre 6-16 segundos, con tamaños variados entre 300x200 pixeles a 600x400 pixeles, con poca variedad de movimiento y poca variedad de luminosidad, con rostros lo más alineados a la cámara, considerada como base de datos de primera generación.

4.2 Procesamiento del Dataset

Se utilizó OpenFace2 para extraer los rostros y se garantizó que todas las imágenes fueran rostros humanos, y posteriormente se creó un tensor el cual se forma con imágenes colocadas en arreglos con valores de entre 0 y 255, y con dimensiones como se muestra en la Tabla 2. Se colocó las etiquetas de dichas imágenes en un arreglo de vectores donde [1,0] identificara rostros falsos, y [0,1] rostros reales o verdaderos.

Tabla 5. Exactitud de los modelos 10% + 10 contra modelos sin AP 20%.

	Modelo 20 %			**Modelo 10% + 10**			**Diferencia**		
	Entrenamiento	Test	Tiempo	Entrenamiento	Test	Tiempo	Entrenamiento	Test	Tiempo
EF0	96.85%	96.77%	7.74'	**97.48%**	**97.31%**	**6.2**'	0.65%	0.55%	19.90%
EF1	**96.60%**	**96.51%**	11.96'	96.26%	96.09%	**9.53**'	-0.35%	-0.44%	20.32%
EF2	96.80%	96.72%	15.39'	**98.13%**	**97.96%**	**11.95**'	1.36%	1.27%	22.35%
EF3	97.31%	97.22%	37.07'	**97.82%**	**97.65%**	**20.84**'	0.52%	0.44%	43.78%

Tabla 6. Exactitud de los modelos AP 20% + 30 contra modelos sin AP 50%.

	Modelo 50 %			**Modelo 20% + 30**			**Diferencia**		
	Entrenamiento	Test	Tiempo	Entrenamiento	Test	Tiempo	Entrenamiento	Test	Tiempo
EF0	**97.33%**	**97.62%**	23.5	96.98%	96.93%	**12.49**	-0.36%	-0.71%	46.85%
EF1	96.99%	**97.67%**	36.82	**97.32%**	97.24%	**19.99**	0.34%	-0.44%	45.71%
EF2	**97.70%**	**97.67%**	47.4	97.56%	97.51%	**24.29**	-0.14%	-0.16%	48.76%
EF3	**97.83%**	**97.80%**	78.95	97.14%	97.10%	**52.94**	-0.71%	-0.72%	32.94%

Tabla 7. Exactitud de los modelos AP 50% + 50 contra modelos sin AP 100%.

	Modelo 100 %			**Modelo 50% + 50**			**Diferencia**		
	Entrenamiento	Test	Tiempo	Entrenamiento	Test	Tiempo	Entrenamiento	Test	Tiempo
EF0	**98.30%**	**98.16%**	50.3	98.16%	98.15%	**41.41**	-0.14%	-0.01%	17.67 %
EF1	**98.38%**	**98.36%**	125.06	97.37%	97.35%	**75.79**	-1.04%	-1.04%	39.40 %
EF2	**95.32%**	**95.16%**	320.91	93.07%	93.06%	**158.47**	-2.42%	-2.26%	50.62 %

Para el entrenamiento de cada modelo, se seleccionó una muestra de imágenes hasta alcanzar una cantidad de imágenes como se puede observar en la Tabla 4 donde el porcentaje corresponde a la totalidad de imágenes de la base de datos UADFV. De esta cantidad de imágenes la mitad son falsas y la mitad son verdaderas para tener un subconjunto balanceado.

Se utilizó clasificación a 5 pliegues utilizando 80% de los datos para entrenamiento y 20% para pruebas.

Para crear los modelos base, para la comparación con AP, se utilizó EfficientNet, desde B0 hasta B3 que se nombraron como EF seguido del número de familia. Cada modelo se creó con pesos aleatorios, utilizando el optimizador Adam, con un total de 15 épocas y con un batch_size[2] de 16.

Dichos valores se seleccionaron por dar la mejor relación exactitud-tiempo en experimentos realizados previamente. Los modelos se definieron con el 10%, 20%, 50% y 100% de los datos.

Para los modelos AP se seleccionaron los tensores correspondientes al 20%, 50% y 100% de datos. Los tensores se particionaron en subtensores A y B con el 10% , 10%; 20%, 30% y 50%, 50% de los datos respectivamente. Para compararlos con los modelos base del 20%, 50% y 100%, paso 1 de la Tabla 3.

Posteriormente se definieron los modelos M con los modelos base de 10%, 20% y 50%, dado que ya estaban entrenados, paso 2 de la Tabla 3.

A continuación, a los modelos M se les agregaron el 10%, 30% y 50% para alcanzar el 20%, 50% y 100% de datos respectivamente de los modelos base. En este proceso se garantizó que los datos agregados fueran los correspondientes a los modelos base asociados.

[2] El número de ejemplos en un batch o lote, es decir conjunto de ejemplos usados en una iteración del entrenamiento del modelo.

Por ejemplo, para el modelo base creado con el 10% de los datos, se le agrego el subtensor B correspondiente para tener los mismos datos que el modelo base del 20% de datos. Al modelo generado se le nombro "Modelo 10% +10". Y de manera semejante para el resto de los modelos, paso 3 de la Tabla 3.

Los resultados de los experimentos realizados se muestran en las Tabla 5, 6 y 7 así como el tiempo de entrenamiento donde el mejor resultado esta resaltado en negritas. El tiempo se muestra con el objetivo de visualizar que este se reduce, lo que hace viable realizar el entrenamiento en equipos personales.

5. Discusión de resultados

De los resultados obtenidos puede rescatarse lo siguiente.

El modelo con la mayor diferencia es EF2 Modelo 50% + 50 con EF2 Modelo 100% el cual difiere en exactitud con un -2.42% en entrenamiento y -2.26 en test. Sin embargo, la diferencia promedio en exactitud entre los modelos sin AP a los modelos con AP es de -0.21% en entrenamiento y -0.32% en test.

Por otro lado, los modelos AP, logran superar siempre en tiempo a los modelos sin AP, logrando una mejora significativa del 25.21% en promedio. Siendo la mejora más pequeña en tiempo de 17.67 % en el modelo EF0 Modelo 50% + 50 respecto a EF0 Modelo 100% y la mejora más grande en tiempo de 64.06% en el modelo EF2 Modelo 50% + 50 respecto de EF2 Modelo 100 %.

6. Conclusiones y trabajo futuro

Con los resultados observados en las Tablas 5, 6 y 7 respecto a la ganancia en tiempo de entrenamiento comparado con la perdida de exactitud, respecto de los resultados de los modelos base, donde la máxima exactitud es 98.38% contra 97.37% del modelo AP, se puede concluir que el AP es una alternativa para entrenar modelos en computadoras personales.

El AP pueden ser utilizado para entrenar modelos en equipos con menor potencia de cómputo, con respecto a los utilizados por los equipos de Seferbekov [6] y Zhao H. et al. [8], dado que al trabajar con subtensores se requiere menor espacio en memoria para el entrenamiento.

Para trabajos futuros se buscarán maneras de alcanzar un equilibrio entre la exactitud, el tiempo de entrenamiento y la potencia de cómputo requerida. De la misma manera se planea expandir los experimentos a la base de datos de segunda generación y utilizar la métrica AUC en conjuntos de datos no balanceados.

Referencias

1. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, vol. 64, pp. 131-148 (2020) doi: 10.1016/j.inffus.2020.06.014
2. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton-Ferrer, C.: The deepfake detection challenge (DFDC) dataset. (2020) doi: 10.48550/arXiv.2006.07397

3. Cellan-Jones, R.: Deepfake videos double in nine months. BBC News (2019) https://www.bbc.com/news/technology-49961089
4. Tolosana, R., Romero-Tapiado, S., Fierrez, J., Vera-Rodriguez, R.: Deepfakes evolution: Analysis of facial regions and fake detection performance. Pattern Recognition, In: ICPR International Workshops and Challenges (ICPR 2021) Lecture Notes in Computer Science, vol. 12665, pp. 442–456 (2021) doi: 10.1007/978-3-030-68821-9_38
5. Baltrusaitis, T., Zadeh, A., Lim, Y., Morency, L. P: OpenFace 2.0: Facial behavior analysis toolkit. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66 (2018) doi: 10.1109/FG.2018.00019
6. Seferbekov, S.: Deepfake detection (DFDC) solution. (2021) https://github.com/selimsef/dfdc_deepfake_challenge
7. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (PMLR), vol. 97, pp 6105–6114 (2019) doi: 10.48550/arXiv.1905.11946
8. Zhao, H., Cui, H., Zhou, W.: 2nd place solution for Kaggle deepfake detection challenge. (2021) https://github.com/cuihaoleo/kaggle-dfdc accesado el 20/08/2021
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001 (2001) doi: 10.1109/CVPR.2001.990517
10. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020) doi: 10.1109/CVPR42600.2020.00327
11. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition *(*2019)

ISSN 1870-4069

Development of a Muscle Fatigue Monitoring Tool Using Myo-Electric Signals and IoT

Marco A. Lopez Oroz, Pedro González-Zamora,
Jesus Pacheco,Víctor H. Benítez

Universidad de Sonora,
Mexico

marco.lopezoroz@gmail.com, {pedro.gonzalez,
jesus.pacheco, victor.benitez}@unison.mx

Abstract. In this project, we develop a system that can monitor the state of a muscle group and estimate muscle fatigue by using EMG sensors and an IoT architecture. The captured data are sent to a remote server for processing and displayed through a web app developed using Node-red. The results are calculated by using previously captured data of bicep EMGs of volunteers. The user interface can display the data being captured in real time as well as the results of previous runs. The purpose of this project is to set up a platform that can be used in the industry for ergonomic purposes, as well as be used in the medical field for monitoring and therapeutic purposes.

Keywords: Myoelectric signals, digital twins, IoT.

1 Introduction

Electromyographic signals (EMGs) are biomedical electric signals emitted by the human body to trigger an organ or muscular tissue to perform an action (tension or contraction).

Since EMG signals are inherently complex and contain a lot of noise (from the measurement equipment, ambient radiation, and even the nervous system itself), using this kind of signal is not simple. There are many ways to "decode" these signals, from statistics to artificial intelligence (AI) algorithms [1].

EMGs have been used to study muscle fatigue in diverse studies [2, 3]. Most of them, use The Median Frequency (MNF), Mean Frequency (MDF), and RMS (Root Main Square) [5–7]as indicators of muscle fatigue.

Thanks to the technological advances related to the measurement of EMG signals, capture devices have become increasingly practical and portable. This allows us to take samples in controlled environments (laboratories) and real-life environments. Data transmission can be performed through the Internet of Things (IoT) architectures, allowing people to analyze and monitor data remotely, and for even better monitoring, technologies such as Digital Twins (DT) can be applied.

A Digital Twin is a virtual representation of an object or the state of an object or process by taking data from sensors and signals. This can help engineers detect

problems even before they happen [8]. In this project, a system that can monitor the state of a muscle group in real time by using EMG sensors and an IoT architecture was implemented.

2 Materials and Methods

Based on the Internet of things technology it is possible to give users access to their EMGs data at any time and anywhere. In other words, users can check through a web application the data sets of EMGs that were captured when they were performing an activity.

For this, a commercial capture system was used to measure the EMGs from volunteers, a local server was configured to store and process the EMGs signals, and a remote server placed at the Amazon Web Services facilities[9] was configured to host a database server, mosquito broker[10] and a Web App. The data was processed locally before being sent to the remote server to be less sensitive to the latency of the network.

The stages of development for this project are as follows:

1. Capture system setup,
2. IoT architecture,
3. Data capture,
4. Processing,
5. Analysis results.

2.1 Capture System Setup

The capture system consists of wireless sensors, a base station, and a local server. The architecture of the capture system is shown in Figure 1. The wireless sensors placed at the skin of the volunteer transmit the measured data to a base station.

The base station sent the data to a local server through a USB port and then, the data are processed to extract features that can be used to determine muscle fatigue.

The local server contains a python application that processes the signal. The model of the python application is shown in Figure 2.

The data collected by EMG sensors are captured in the time domain, however in this study was necessary to transform the data to the frequency domain. For this, the python application implements the Fast Fourier Transform [11].

Once the data are in the frequency domain, features such as RMS, MDF, and MNF are extracted and stored in the local server. Subsequently, the stored data are sent to a remote server located in the cloud of Amazon Web Services.

The MNF is obtained following the next steps:

1. The EMG data is transformed into the frequency domain using the Fast Fourier Transform (FT). The FT algorithm used in this work is part of the NumPy library of Python.
2. The Power Spectrum Density (PSD) is calculated by multiplying the previously calculated FT array by its conjugate.

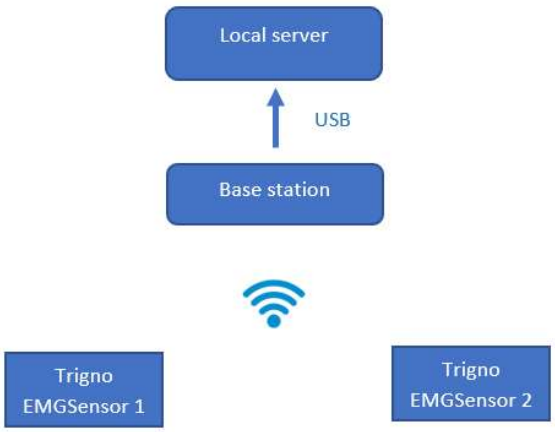


Fig. 1. Capture system setup used in this work. The EMGs signals were measured using wireless sensors that send data to a base station. Then, the base station sends the data to a local server which processes them.

3. Then, the sum of the product of the frequency value of PSD at each frequency bin as well as the sum of the PSD is computed. Finally, the MNF is obtained by dividing the first sum by the second one.

The MDF is the frequency that divides the spectrum into two regions with an equal sum of amplitude. The step 1 and 2 are the same as the ones used to compute the MNF. A more detailed procedure is shown below:

1. The sum of all bins of PSD array is computed and divided by two.
2. The algorithm loop through each frequency bin and sum its PSD value and stop until this sum is equal to half of the total sum of PSD. The frequency where the loop is halted is the median frequency.

The RMS is computed as follows:

1. Given an array of data of length N, each value of the array is squared and then summed.
2. This sum is then divided by N.
3. Finally, the square root of the resulting quotient is computed.

2.2 IoT architecture

The architecture used in this work is based on the 3 layers described in [12]. The setup of the platform implemented in this work is described as follows for each layer.

1. **The perception Layer** refers to the EMGs sensors and the devices used for the captured data. The perception layer represents the capture system setup described

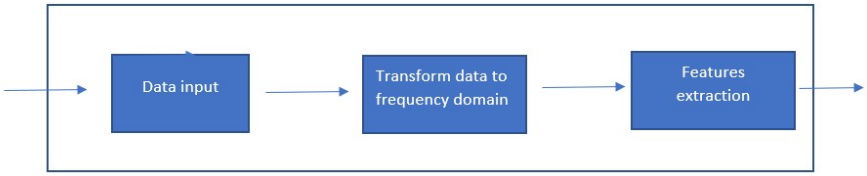


Fig. 2. Data processing using a python application.

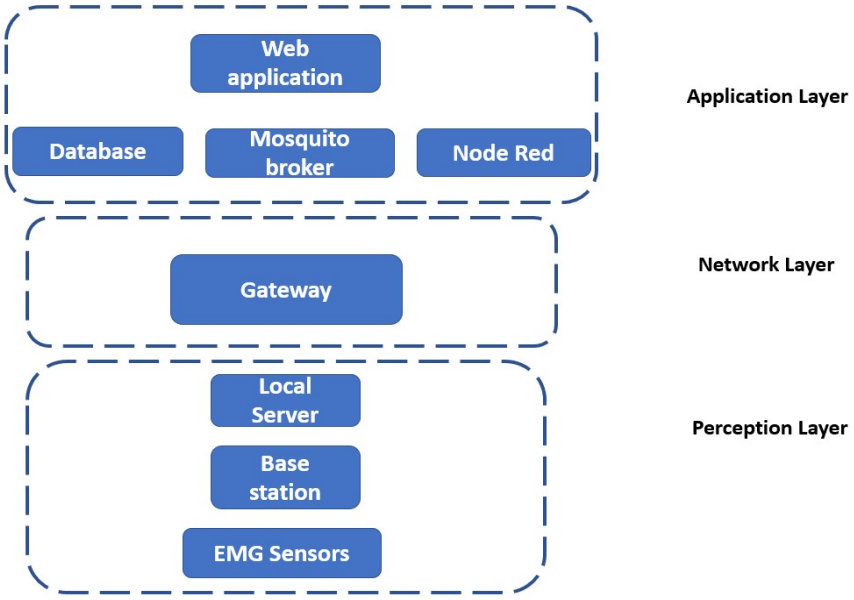


Fig. 3. IoT architecture used in this work.

in the previous section. The EMGs data are captured at frequencies around the 2000 HZ, so that, this will require a high compute performance if the Web App shows in real-time the variables related to muscle fatigue as MNF, RMS, and MDF. For this reason, in this work, only EMG raw data is shown live through the Web App and the muscle fatigue variables are computed offline in a local server located at this layer and sent to the remote server after finishing the data taking.

2. **Network Layer** is the one that describes all the devices that make possible the communication between the Perception Layer and the Application Layer. So, it includes elements such as routers, wireless devices, and the protocols used for the data interchange. The MQTT protocol [10] was used to send the EMGs raw data to the Node-Red platform [1] locate in the cloud. The preprocessed data is sent to the database using the FTP protocol.
3. **Application Layer** is located at the Cloud Server. In this work, Amazon Web Services is used to host a remote no SQL database, a mosquito broker, and a Node Red server. Moreover, the Node Red Server implements a Web App where information related to previous runs of a specific user can be accessed.

Table 1. Body composition of the volunteer that participated in this experiment.

Id	Weight	Body fat	Muscle mass
1	111.3	52.5	21.6
2	59.4	31.6	29.1
3	89.1	47.5	23.5

Fig. 4. Activity performed during data capture.

2.3 Data Apture

Data capture was done using volunteer students from the University of Sonora. A sensor was attached to their bicep and they were asked to perform standing bicep curls (Figure 4) with different weights to analyze the impact of different loads on the myoelectric signals. In this case, 5lb and 10lb dumbbells were used. The body composition of the volunteers is shown in Table 1.

2.4 Processing

The captured data is saved into .csv files and then the features are extracted. The computed features are Root Mean Square (RMS), Mean Frequency (MNF), and Median Frequency (MDF).

The features were chosen as the literature shows a clear relationship between them and muscle fatigue. As fatigue increases, RMS increases, and frequencies decrease (Figure 5). The RMS, MNF, and MDF can be checked offline through the Web App as it is shown in Figure 6. Users can access their information after the data capture as the data are stored in the remote database. The user only has to log in to the platform and select the data sample.

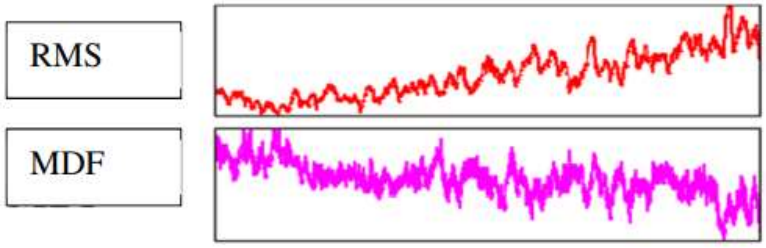


Fig. 5. Expected behavior of RMS and MDF when muscle fatigue is becoming higher [14].

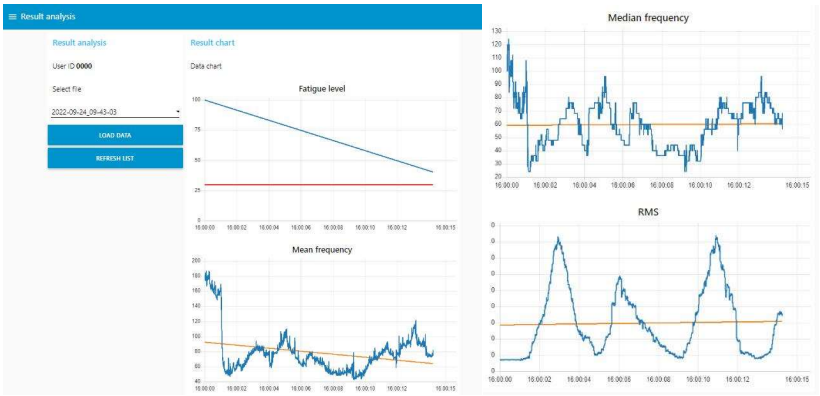


Fig. 6. Web app interface of the platform developed in this work.

3 Results

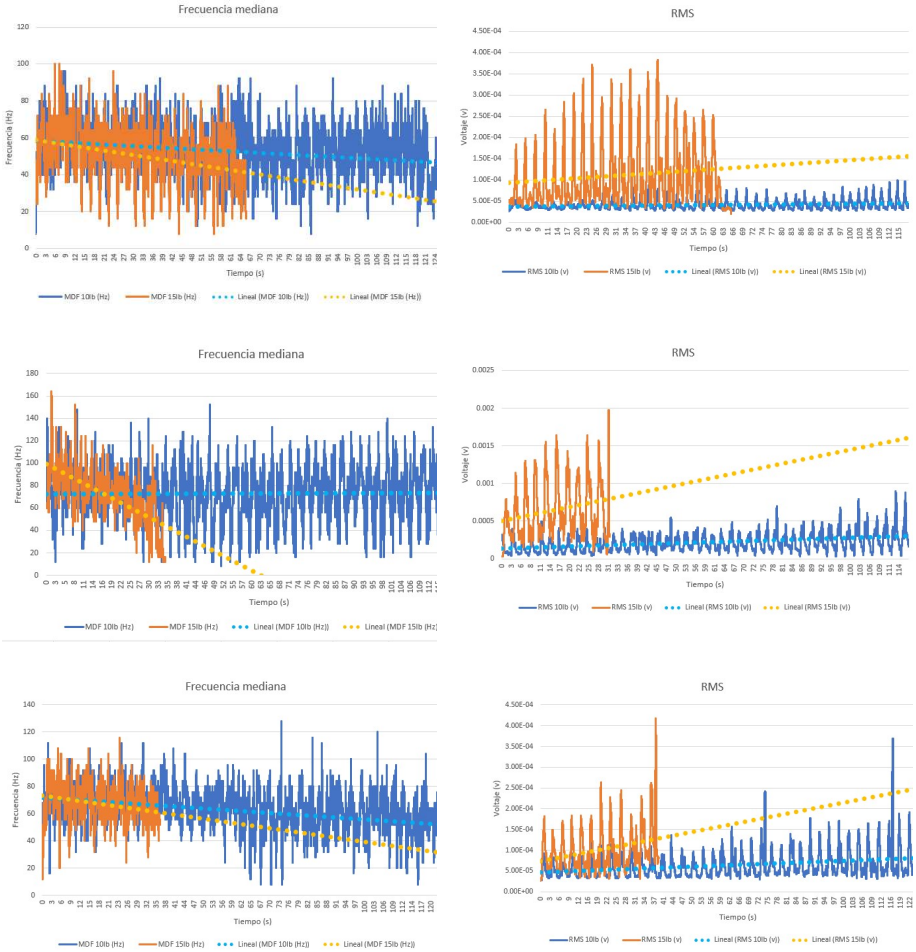
The results obtained (so far) are shown in figure 7. The results shown in Figure 7 follow a similar behavior to the ones shown in Figure 5 as expected. As fatigue increases, RMS increases, and frequency decreases.

The results also show that the average slope is more critical when a higher load is applied to the muscle group. The above results were stored for each user in the created platform and each user can.

4 Conclusions and Future Work

A platform to capture and process EMGs data was developed in this work. The system can measure myoelectric signals from users and store them in a remote server after being processed. This brings the capability to the system to extract features related to muscle fatigue. Moreover, the platform architecture is based on an IoT architecture to give continuous access to users to check the data anytime and anywhere.

The tool can be used in the industry to monitor the status of muscle fatigue of their employees and it can be also used in the field o medicine.



(a) (b)

Fig. 7. Typical MNF (a) and RMS (b) behavior of 3 volunteers.

The future work is to use the variables related to muscle fatigue as RMS, MNF, and MDF to give valuable information that allows people to make strategies to create better working environments.

References

1. Yousif, H. A., Zakaria, A., Rahim, N. A., Salleh, A. F. Bin, S., Mahmood, M., Alfarhan, K. A., Kamarudin, L. M., Mamduh, S. M., Hasan, A. M., Hussain, M. K.: Assessment of muscles fatigue based on surface EMG signals using machine learning and statistical approaches: A review. In: IOP Conference Series: Materials Science and Engineering vol. 705, no. 012010 (2019) doi: 10.1088/1757-899X/705/1/012010

2. Cifrek, M., Medved, V., Tonković, S., Ostojić, S.: Surface EMG based muscle fatigue evaluation in biomechanics. Clinical Biomechanics, vol. 24, no. 4, pp. 327–340 (2009) doi: 10.1016/j.clinbiomech.2009.01.010
3. Reaz, M. B., Hussain, M. S., Mohd-Yasin, F.: Techniques of EMG signal analysis: Detection, processing, classification and applications. Biological Procedures Online, vol. 8, pp. 11–35 (2006) doi: 10.1251/bpo115
4. Phinyomark, A., Thongpanja, S., Hu, H., Phukpattaranont, P., Limsakul, C.: The usefulness of mean and median frequencies in electromyography analysis. Computational Intelligence in Electromyography Analysis, A Perspective on Current Applications and Future Challenges (2012) doi: 10.5772/50639
5. Wang, L., Wang, Y., Ma, A., Ma, G., Ye, Y., Li, R., Lu, T.: A comparative study of EMG indices in muscle fatigue evaluation based on grey relational analysis during all-out cycling exercise. BioMed Research Internatinal, pp. 1–8 (2018) doi: 10.1155/2018/9341215
6. Zhang, G., Morin, E., Zhang, Y., Etemad, S. A.: Non-invasive detection of low-level muscle fatigue using surface EMG with wavelet decomposition. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp. 5648–5651 (2018) doi: 10.1109/EMBC.2018.8513588
7. Fuller, A., Fan, Z., Day, C., Barlow, C.: Digital twin: Enabling technologies, challenges and open research. IEEE Access, vol. 8, pp. 108952–108971, (2020) doi: 10.1109/ACCESS.2020.2998358
8. Amazon web services: Overview of Amazon web services AWS whitepaper (2022)
9. Shahri, E., Pedreiras, P., Almeida, L.: Extending MQTT with real-time communication services based on SDN. Sensors, vol. 22, no. 9 (2022) doi: 10.3390/s22093162
10. Brigham, E. O., Morrow, R. E.: The fast Fourier transform. IEEE Spectrum, vol. 4, pp. 63–70 (1967) doi: 10.1109/MSPEC.1967.5217220
11. Lombardi, M., Pascale, F., Santaniello, D.: Internet of things: A general overview between architectures, protocols and applications. Information, vol. 12, no. 2 (2021) doi: 10.3390/info12020087
12. Lekic, M., Gardasevic, G.: IoT sensor integration to Node-RED platform. In: 2018 17th International Symposium, INFOTEH, pp. 1–5 (2018). doi: 10.1109/INFOTEH.2018.8345544

ISSN 1870-4069

Desarrollo de un sistema embebido para advertir sobre las condiciones de riesgo de contagio de COVID-19 mediante el monitoreo de la calidad del aire

Yair Romero López, Ricardo Álvarez González,
Rodrigo Lucio Maya Ramírez, Alba Maribel Sánchez Gálvez

Benemérita Universidad Autónoma de Puebla,
México

yair.romerolope@alumno.buap.mx, {ricardo.alvarez, rodrigo.maya,
alba.sanchez}@correo.buap.mx

Resumen. El impacto súbito y generalizado de la pandemia del COVID-19, ha afectado a las actividades esenciales y presenciales de nuestra vida cotidiana, a pesar de la existencia de nuevas variantes del virus, existe un número significativo de la población vacunada en México, esto ha dado origen a la propuesta del retorno a las actividades presenciales, en sectores educativos, turísticos, profesionales, etc. Por ello se requiere todas las medidas de protección e higiene, además de una planificación muy estricta, para minimizar el riesgo de contagio. La medición de la concentración de dióxido de carbono CO_2 es una estrategia que puede advertir el riesgo de contagio de la enfermedad del coronavirus (COVID-19) en un espacio cerrado donde se encuentre reunido un grupo de personas. El resultado puede proporcionar información, a partir de la cual se puede deducir sí la ventilación es adecuada o deficiente, lo que facilitaría la propagación del virus. Esta es la razón por la cual se propone desarrollar un sistema embebido para monitorear la concentración de CO_2 en un espacio cerrado y emitir alarmas cuando se exceda el valor de 1000 ppm y adicionalmente generar un registro diario de los datos, almacenándolos en la nube para su posterior análisis.

Palabras clave: COVID-19, CO_2, sistema embebido, IoT, sensor.

Development of an Embedded System for Warning about COVID-19 Contagion Risk Conditions through Air Quality Monitoring

Abstract. The sudden and widespread impact of the COVID-19 pandemic has affected the essential and face-to-face activities of our daily lives, despite the existence of new variants of the virus, there is a significant number of the population vaccinated in Mexico, this has given rise to the proposal to return to face-to-face activities, in educational, tourist, professional sectors, etc. For this reason, all protection and hygiene measures are required, in addition to very strict planning, to minimize the risk of contagion. The measurement of the concentration of carbon dioxide CO_2 is a strategy that can warn of the risk of

contagion of the coronavirus disease (COVID-19) in a closed space where a group of people is gathered. The result can provide information, from which it can be deduced whether ventilation is adequate or poor, which would facilitate the spread of the virus. This is the reason why it is proposed to develop an embedded system to monitor the concentration of $C0_2$ in a closed space and issue alarms when the value of 1000 ppm is exceeded and additionally generate a daily record of the data, storing it in the cloud for further analysis.

Keywords: COVID-19, $C0_2$, embedded system, IoT, sensor.

1 Introducción

En los últimos años, la calidad del aire en interiores ha recibido una atención considerable de gobiernos ambientales, instituciones políticas y la comunidad de científicos internacionales, debido a su estrecha asociación con la salud pública, la comodidad y el bienestar de las personas [1, 2].

Actualmente se puede hacer uso de las tecnologías emergentes para advertir sobre las condiciones de riesgo de contagio del COVID-19, el Internet de las cosas (IoT) muestra un gran potencial para medir datos en tiempo real de un ambiente cerrado que puedan auxiliar a los ocupantes de éstos espacios a tomar decisiones relevantes para tomar medidas necesarias de ventilación y así mitigar los niveles de $C0_2$ y de esta manera evitar una diseminación del virus en los ocupantes [3].

El Internet de las cosas es un tema ampliamente discutido en el campo de las tecnologías de la comunicación. Esta tecnología puede conectar una cantidad ilimitada de dispositivos y sensores para influir en la forma en que trabajamos y vivimos. IoT encuentra múltiples aplicaciones en escenarios del mundo real mientras automatiza procesos para aumentar la productividad [2].

La información registrada a través de un sensor $C0_2$ en ambientes cerrados y públicos puede proporcionarnos valiosa información, puede ser vista por expertos en salud pública, funcionarios gubernamentales y legisladores, para tomar medidas de mejora en la salud y el bienestar social [3].

2 Propósito y enfoque del proyecto en el combate de los contagios de COVID -19 en espacios cerrados

Actualmente la OMS plantea que el COVID-19, se transmite principalmente a través de microgotas. La inoculación de microgotas en las vías respiratorias deviene de la exposición del hospedador a eventos del paciente (tos, estornudo, carraspeo, etc.) o procedimientos que inducen dispersión de gotitas en el aire.

Para la OMS la mayoría de los contagios se producen a través del contacto cercano, motivo por el cual el distanciamiento entre personas debe ser mínimo de un metro y una ventilación adecuada en espacios cerrados es primordial [4, 5].

El monitoreo de $C0_2$ es un medio establecido para evaluar si la ventilación es adecuada para el número de personas que ocupan un espacio cerrado. Aunque los niveles de $C0_2$ oscilan entre 350 y 450 ppm al aire libre, las personas que se reúnen y respiran dentro de un edificio harán que el $C0_2$ se acumule a niveles mucho más altos a

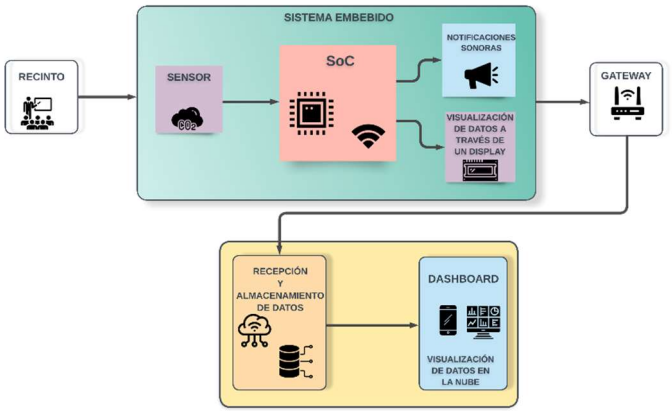


Fig. 1. Diagrama general del Sistema.

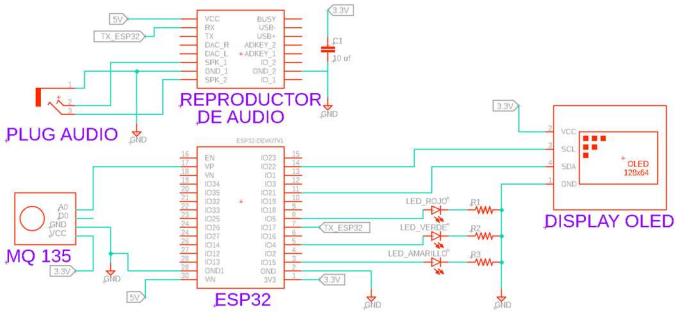


Fig. 2. Diagrama eléctrico del sistema.

menos que se elimine a través de la ventilación. Cuantas más personas ocupen un espacio y más intensa sea su actividad física, más ventilación se requiere para mantener la comodidad de los ocupantes. Se considera que arriba de 1000 ppm del $C0_2$ en un espacio cerrado podría ayudar a facilitar a una transmisión de COVID-19 [6].

A pesar de existir nuevas variantes del virus, el medio de propagación es el mismo, por lo cual la detección de niveles elevados de $C0_2$ se puede emplear cómo método efectivo de riesgo de transmisión de cualquier cepa del virus [5].

Además, la cantidad de personas y la naturaleza de sus actividades dentro de un espacio pueden afectar la calidad del aire y la presencia de partículas infecciosas a lo largo del tiempo. En este trabajo se propone medir las concentraciones de $C0_2$ en interiores utilizando sensores de bajo costo como una forma de estimar el riesgo de transmisión de COVID-19 en interiores [7].

3 Estado del arte

En el mercado encontramos una amplia variedad de sistemas de monitoreo de la calidad el aire sin embargo varios sistemas de monitoreo basados en IoT carecen de la gene- ración de notificaciones sonoras de voz para alertar a los usuarios finales sobre

los niveles óptimos de convivencia y sin riesgo de un contagio masivo en el lugar a monitorizar [8]. Los dispositivos que cuentan con conectividad a la nube tienen costos altos o incluso para algunos es necesario adquirir algún complemento para la conexión a internet.

Actualmente en la literatura se destacan diferentes sistemas embebidos relacionados a la calidad del aire en espacios cerrados, un ejemplo del desarrollo de estos sistemas es el propuesto el en articulo [9], en el cual se presenta el diseño e implementación de un sistema de monitoreo de la calidad del aire mediante el sensado del $C0_2$, se puede apreciar que no se obtiene un registro de las mediciones realizadas y carece de notificaciones sonoras de voz al existir un riesgo al presentarse un nivel alto de partículas de $C0_2$ en el ambiente.

A nivel comercial existen diversos productos en el mercado, donde los monitores de $C0_2$ se conectan a la nube para el envío de datos, un ejemplo de ello es el sistema comercial [10], este sistema cuenta con una aplicación móvil para observar los datos adquiridos del monitor, además de contar con una pantalla en el monitor para poder visualizar los datos, la aplicación cuenta con registros de las mediciones obtenidas de días anteriores, sin embargo se observa que carecen de alarmas para advertir las condiciones de riesgo de contagio de COVID-19 a las personas que se encuentren en el recinto.

La mayoría de los sistemas desarrollados comercialmente y con bajos costos, no siguen los procedimientos de calibración adecuados ni las pruebas de confiabilidad antes del despliegue de los sensores IAQ (Indoor air quality) en las ubicaciones objetivo [8].

4 Desarrollo

Se desarrolla un sistema embebido para monitorear la concentración de $C0_2$ en un espacio cerrado y así emitir alarmas sonoras de voz cuando se excedan los parámetros establecidos, adicionalmente se lleva unregistro diario de los datos, almacenándolos en la nube para su posterior análisis.

El proyecto está dividido en dos bloques fundamentales mostrados en la figura 1 los cuales se describen a continuación:

- Sistema embebido.
- Recepción, almacenamiento y visualización de de los datos obtenidos en la nube.

4.1 Sistema embebido

El diagrama eléctrico del sistema embebido mostrado en la Figura 2 está conformado por cuatro bloques fundamentales:

- **Sensor de $C0_2$:** Módulo que incluye un sensor de gas el cual detecta las partes por millón del $C0_2$ en el ambiente, en este caso se hace uso del MQ135 debido a su bajo costo y sus altas prestaciones para la detección de partículas nocivas para el ser humano, este sensor cuenta con la capacidad de detección de 10-1000 ppm ppm de $C0_2$ [11].

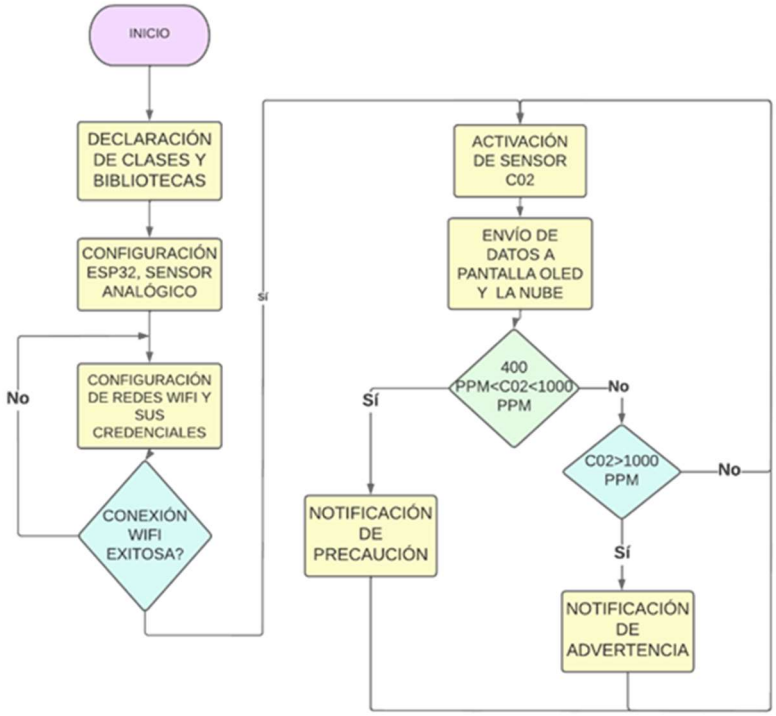


Fig. 3. Algoritmo implementado en el módulo electrónico ESP32.

- **Tarjeta de desarrollo ESP 32:** Tarjeta de control electrónica que contiene el SoC (System on a Chip) y el firmware correspondiente, al igual que la respectiva etapa electrónica para él envió de información inalámbrica vía WiFi, este SoC es una excelente opción para el desarrollo de sistemas embebidos con conectividad IoT ya que el soporte de la comunidad científica y académica se ha centrado en este dispositivo que se dirige a una amplia variedad de aplicaciones industriales y médicas [12, 13].
- **Display electrónico:** Dispositivo el cual notifica de forma visual los datos obtenidos a través del dispositivo que exhibe el resultado en una pantalla de tecnología OLED SSD130 ya que tiene ventaja de tener un consumo bajo debido a que solo se enciende el pixel necesario y no requieren de luz de fondo y tienen una mejor visibilidad en ambientes luminosos, como bajo el sol en comparaciones de otros exhibidores [14].
- **Notificaciones sonoras:** Adicionalmente el sistema embebido cuenta con notificaciones mediante voz, para advertir las condiciones en el cual se encuentra la calidad del aire y el riesgo que implica, se hace uso de un DFplayer mini reproductor mp3 con salida de audio a unas bocinas con plug 3.5 mm, este dispositivo fue seleccionado ya que es un módulo electrónico de fácil implementación y con la capacidad de reproducción de pistas de audio en los formatos más comunes, cómo lo son MP3, WMA y WAV [15].

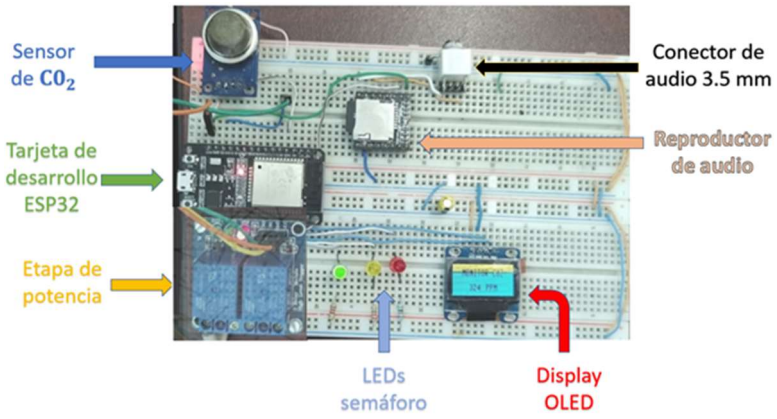


Fig. 6. Sistema embebido desarrollado en una placa de prototipado electrónico.

4.2 Módulo electrónico ESP32

Actualmente el SoC de ESP32 es una excelente opción para el desarrollo de sistemas que van desde redes de sensores de bajo consumo hasta las tareas más exigentes, como la codificación de voz, la transmisión de música y la decodificación de MP3, redes de sensores industriales, etc. [13].

Además de contar con conectividad inalámbrica Bluetooth y WiFi en un solo módulo. De igual manera el costo de estos dispositivos es accesible y sus funcionalidades son amplias para el desarrollo de proyectos de IoT con prestaciones de bajo consumo [8].

4.3 Algoritmo implementado para el envío de datos

El envío de datos llevado a cabo por el ESP32 se realiza mediante la creación de un algoritmo con el lenguaje de programación Python haciendo uso del interprete MicroPhyton, esto debido a la fácil integración de protocolos de comunicación usados en redes IoT [16]. En la figura 3 podemos apreciar el algoritmo programado en MicroPhyton.

4.4 Envío de datos hacienda uso del protocolo de comunicación MQTT

Se han desarrollado varios protocolos y métodos para la comunicación de datos en el campo de IoT. En el presente trabajo se utilizará el protocolo MQTT (Message Queue Telemetry Transport), el modelo es mostrado en la figura 4, basado en una comunicación de publicación-suscripción.

El componente que produce la información pública la información, por lo que se le denomina *Publisher*, el cual es nuestro caso es el sistema embebido. El interesado en recibir la información publicada se denomina suscriptor.

El *Broker* se asegura de que el suscriptor reciba todos los datos que se han publicado sin ninguna pérdida, en nuestro caso la aplicación IoT denominada *ThingSpeak*. MQTT es un sistema de publicación y suscripción basado en temas denominados *Topics*. Esto

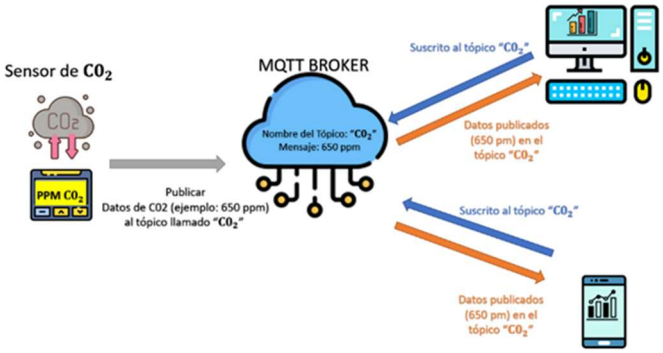


Fig. 4. Modelo de comunicación empleado.

Fig. 5. *Dashboard* generada en el software *ThingSpeak.*

significa que el *Publisher* produce la información sobre un *Topic* que puede ser suscrito por cualquier dispositivo suscriptor [17].

4.5 Recepción, almacenamiento y visualización de los datos obtenidos

Se hace uso de la plataforma IoT *ThingSpeak*, la cual se conecta con el sistema embebido mediante un protocolo de comunicación inalámbrica MQTT [17], de este modo, se llevará a cabo la recepción y almacenamiento de los datos obtenidos. Se podrá descargar las distintas mediciones llevadas a cabo en un archivo con extensión .CSV.

Los datos obtenidos se presentan a través del *dashboard* mostrado en la figura 5, indicando de manera gráfica las mediciones de CO_2 realizadas.

5 Pruebas y resultados

Para corroborar el correcto funcionamiento del sistema se realizaron pruebas con el prototipo mostrado en la figura, el cual se colocó en un aula de clases con las siguientes dimensiones, 6 x 4 x 2.5 m:

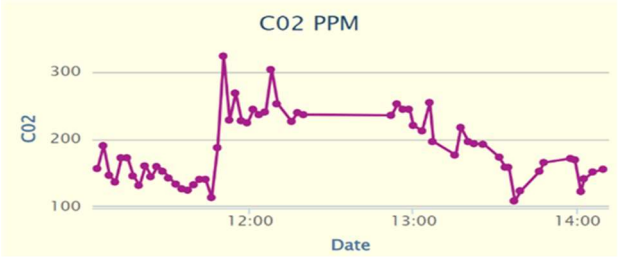


Fig. 7. Gráfica mostrada en el software *ThingSpeak*.

Tabla 1. Formato de la tabla generada a partir de las mediciones obtenidas.

Fecha	PPM Sensadas
2022-06-24 16:48:10 UTC	187
2022-06-24 16:50:23 UTC	323
2022-06-24 16:52:35 UTC	350
2022-06-24 16:54:48 UTC	396
2022-06-24 16:57:00 UTC	410
2022-06-24 16:59:12 UTC	412
2022-06-24 17:01:24 UTC	421
2022-06-24 17:03:37 UTC	408
2022-06-24 17:05:48 UTC	400

- Aula con ventanas y puertas abiertas, una ventilación de aire considerablemente buena y sin ocupantes.
- Aula con ventanas y puertas cerradas, una ventilación de aire considerablemente mala y con cinco ocupantes.
- Aula con ventanas y puertas abiertas, una ventilación de aire considerablemente buena y con cinco ocupantes.

Los resultados obtenidos y graficados en la dashboard del software *ThingSpeak* son mostrados en la figura 7, en las primeras horas de 11:00 a 12:00 existe una buena ventilación y no hay ocupantes por lo cual los niveles de $C0_2$ son bajos.

Posteriormente se observa una alza de las mediciones de ppm del $C0_2$ ya que la ventilación era deficiente y existían cinco ocupantes en el aula.

Finalmente se mantienen los mismos cinco ocupantes pero se ventila el recinto y las ppm del $C0_2$ bajan gradualmente.

5.1 Notificaciones de audio

El sistema embebido contiene un conector hembra de 3.5 mm, conectado a un módulo reproductor de audio, donde se reproducen pistas en formato MP3. Se cuentan con dos diferentes audios para indicar los siguientes niveles de calidad del aire y el posible riesgo existente de un contagio de COVID-19:

- $400 < CO_2 <= 1000$ PPM : ¡Precaución!, la calidad del aire empieza a ser crítica, por la seguridad de los asistentes considere ventilar el recinto y minimizar el número de ocupantes.
- $CO_2 > 1000$ PPM : ¡Advertencia!, Ambiente no óptimo para la convivencia, el dióxido de carbono sobrepasa los niveles de riesgo, favor de desalojar y ventilar el recinto, gracias.

5.2 Tabla de datos obtenidos

Los datos se almacenarán en la cuenta enlazada al software de *ThingSpeak*, donde tendremos acceso a la dashboard y a una tabla de datos con la información obtenida del sensor, se puede observar en la Tabla 1 que $CO_2 > 400$ ppm en este momento se activa la notificación sonora de precaución lo cual le indica a los ocupantes ventilar el área para reducir los niveles de CO_2, posteriormente se observa una reducción de las ppm. Se pueden visualizar los valores obtenidos por el sensor en la Tabla 1 dicha información es almacenada en un archivo con formato .CSV.

6 Conclusiones

El sistema en esta etapa de desarrollo funcionó satisfactoriamente ya que cuenta con una conexión a internet del sistema embebido con una base de datos en donde se almacena la información adquirida del sensor de CO_2, de esta manera se obtiene un parámetro para evaluar un posible riesgo mayor de contagio de COVID-19, esto será benéfico para trabajos en el futuro en donde se haga una trazabilidad del comportamiento de los niveles de CO_2 y su correlación con algunos factores, cómo lo son el clima, la hora, el número de ocupantes, etc.

El sistema embebido diseñado puede ser implementado en lugares cerrados, por ejemplo, en aulas de clase o en oficinas de edificios públicos donde la generación de notificaciones sonoras de voz pueda advertir sobre un nivel de riesgo de manera oportuna a los ocupantes del lugar.

Referencias

1. Organización panamericana de la salud (2022) www.paho.org/es/temas/calidad-aire
2. Sultana, S.: A comparison study of air pollution detection using image processing, machine learning and deep learning approach. Global journal of computer science and technology, vol. 19, no. 1 (2019)
3. Saini, J., Dutta, M., Marques, G.: Internet of things for indoor air quality monitoring. Springer International Publishing (2021) doi: 10.1007/978-3-030-82216-3
4. Bejarano¸ D.: Modos de transmisión y diseminación interhumana del virus SARS-CoV-2. Revista de salud publica del Paraguay, vol. 1, no. 11 (2021)
5. Organización mundial de la salud (2022) www.who.int/es/activities/tracking-SARS-CoV-2-variants
6. O´Keeffe, J.: Air cleaning technologies for indoor spaces during the COVID-19 pandemic (2020) ncceh.ca/content/blog/air-cleaning-technologies-indoor-spaces-during-covid-19- pandemic

7. United States environmental protection agency.: Particulate Matter (PM) Pollution (2022) www.epa.gov/pm-pollution/particulate-matter-pm-basics
8. Peng, Z., Jimenez, J. L.: Exhaled Co2 as a COVID-19 infection risk proxy for different indoor environments and activities. Environmental Science and Technology Letters, vol. 8, no. 5, pp. 392–397 (2021) doi: 10.1021/acs.estlett.1c00183
9. Cordero-Picadp, K., Monge-Sanabria, R. N., Segura-Vargas, C. D., Morales-Hernández, S.: Diseño interdisciplinario de un sistema de sensado de CO2 para enfrentar la pandemia en los espacios cerrados del Tecnológico de Costa Rica (2022) https://doi.org/10.18845/tm.v35i5.6189
10. Domodesk (2022) www.domodesk.com/1520-medidor-co2-y-calidad-del-aire-con-8-funciones-wifi.html
11. Technical data MQ-135 gas sensor (2022) pdf1.alldatasheet.com/datasheetpdf/view/1307647/WINSEN/MQ135.html
12. Beningo, J.: Cómo seleccionar y usar el módulo ESP32 con Wi-Fi/Bluetooth adecuado para una aplicación de IoT industrial (2022) www.digikey.com.mx/es/articles/how-to-select-and-use-the-right-esp32-wi-fi-bluetooth-module
13. Espressif Systems. Esp32 serie (2022) www.espressif.com/sites/default/files/documentation/ esp32_datasheet_en.pdf
14. DFPLayer Mini (2022) picaxe.com/docs/spe033.pdf
15. SSD1306 Technical data (2022) cdn-shop.adafruit.com/datasheets/SSD1306.pdf
16. Python for microcontrollers (2022) micropython.org
17. Nettikadan, D., Subodh-Raj, M. S.: Smart community monitoring system using thingspeak Iot plaform. International Journal of Applied Engineering Research, vol. 13, pp. 13402–13408 (2018)

ISSN 1870-4069

Diseño de un sistema de gestión de datos climáticos bajo una metodología de desarrollo PSP

Juan Pablo Báez-Vásquez, Alberto Portilla-Flores

Universidad Autónoma de Tlaxcala,
Facultad de Ciencias Básicas, Ingeniería y Tecnología,
México

{jpablob@msn.com, alberto.portilla.f}@uatx.mx

Resumen. Los sistemas de observación del clima concentran una gran cantidad de información obtenida a partir de una gran variedad de fuentes como las redes de estaciones meteorológicas, previo a su aprovechamiento es preciso verificar la calidad de los datos; un Sistema de Gestión de Datos Climáticos (CDMS) es la herramienta para la recopilación, gestión y verificación de esos datos. En este artículo se propone un diseño basado en las recomendaciones a cumplir por un CDMS y se propone el uso de una metodología PSP (Personal Software Process) para el desarrollo del mismo.

Palabras clave: Gestión de datos, datos climáticos, ingeniería de SW, PSP, sistemas distribuidos, EMA.

Design of a Climatic Data Management System Under a PSP Development Methodology

Abstract. Climate observation systems concentrate a large amount of information obtained from a wide variety of sources such as weather station networks. Prior to using it, it is necessary to verify the quality of the data; a Climate Data Management System (CDMS) is the tool for the collection, management and verification of these data. This article proposes a design based on the recommendations to be met by a CDMS and the use of a PSP (Personal Software Process) methodology for its development is proposed.

Keywords: Data management, climate data, SW engineering, PSP, distributed systems, EMA.

1. Introducción

El clima se define como las condiciones promedio del sistema climático de un lugar y en un periodo de tiempo concreto, determinado por la interacción de los componentes del sistema (atmósfera, litósfera, criosfera, hidrósfera y biósfera) que es influenciado por la radiación solar y las características de cada elemento; el sistema climático precisa

las condiciones promedio de variables como la temperatura y la precipitación [1]. La climatología aplicada emplea la información y conocimientos meteorológicos y climatológicos para resolver y atender problemas sociales, económicos y medioambientales; por ejemplo, el cambio climático, reducción de riesgo de desastres, desarrollo urbano sustentable, etc.

De lo anterior se deduce la conveniencia de disponer de información exacta y oportuna [2], dicha información derivada de datos meteorológicos y climáticos se sustenta en datos recopilados mediante la observación y el registro amplio y sistemático del clima. El Sistema Mundial de Observación del Clima proporciona observaciones del estado de la atmósfera y la superficie de los océanos [3], los dispositivos empleados con ese fin generalmente están a cargo de los Servicios Meteorológicos e Hidrológicos Nacionales (SMHN) entre otras organizaciones.

Tales dispositivos se agrupan en lo que se denomina redes de estaciones meteorológicas, utilizadas para realizar desde pronósticos agrometeorológicos hasta proyecciones de clima futuro; estas redes generan un gran volumen de datos que deben ser gestionados adecuadamente. Entre los dispositivos que conforman las redes de observación están las Estaciones Meteorológicas Automáticas (EMA), estas no requieren de personal técnico para realizar las observaciones.

Sin embargo, los elementos que conforman una red de estaciones no están exentos de desperfectos o errores en las mediciones; por ello se consideran como fundamentales las actividades rutinarias de verificación de calidad y la homogeneidad de los datos recabados; los sistemas informáticos favorecen la realización de estas tareas a la vez que optimizan su consulta e interpretación.

En este sentido, un Sistema de Gestión de Datos Climáticos (Climate Data Management System o CDMS) es una valiosa herramienta de apoyo para los administradores de las redes meteorológicas que agilizan el trabajo relacionado con la verificación, almacenamiento y difusión de los datos.

La Organización Meteorológica Mundial (OMM) ha publicado distintos documentos con la intención de servir como guías en el desarrollo de los CDMS, se trata de un marco o modelo que describe de forma abstracta los componentes requeridos para la funcionalidad del sistema [4], aunque no definen ni especifican las técnicas o metodologías de la ingeniería de software (SW) que se deben emplear para el desarrollo del CDMS y en consecuencia cumpla y cubra las expectativas y requerimientos de los usuarios de los datos climáticos.

Este sistema que se llevará a cabo de forma individual a priori puede describirse como un desarrollo de software personal y precisamente es que PSP (Personal Software Process) provee las herramientas para plantearlo, afrontarlo y llevar un control adecuado del proyecto mejorando además los métodos de trabajo del ingeniero de SW.

2. Planteamiento del problema

El Departamento de Investigaciones Arquitectónicas y Urbanísticas (DIAU) administra una red de estaciones meteorológicas denominada Red Automática de Monitoreo Meteorológico (RAMM), sin embargo, no cuenta con un sistema informático de apoyo, es decir, no se dispone de herramientas automatizadas para realizar la inspección y control de calidad de los datos; tampoco de una interfaz que

facilite la consulta ni la extracción de la información para su empleo en proyectos de investigación.

Algunas de las causas de ausencia de calidad y de homogeneidad más comunes pueden ser resultado de fallos atribuibles a los instrumentos por el deterioro de sus componentes como los valores fuera del rango de medición, los procesos de transmisión de datos, cambios efectuados durante el mantenimiento de una estación o el cambio de su ubicación, la introducción de datos mediante teclado y su validación.

Cuando se presenta la necesidad de procesar un conjunto de datos se lleva a cabo con el apoyo de hojas de cálculo o herramientas especializadas, por ejemplo, para completar series de datos y para detectar valores fuera de rango, todas ellas requieren de archivos con formatos y estructuras particulares. Se debe considerar también el volumen de los datos, cada EMA acumula el promedio de las observaciones de cada 15 minutos, esto es 96 promedios diarios, lo que representan 595,680 registros por año en conjunto por las 17 estaciones de esta red.

Después de realizar los procesos de verificación de los datos estos se deben conservar, pero actualmente tampoco se cuenta con las herramientas para resguardarlos en una base de datos (BD) de información procesada, ni mucho menos para generar las estadísticas y gráficas a partir de ellos; por tanto, es relevante e indispensable que en un corto plazo se desarrolle e implemente un sistema para el procesamiento de los registros obtenidos por la RAMM de lo contrario con el tiempo representará una tarea sustancialmente difícil de manejar.

3. Especificaciones de los CDMS

Las guías de la OMM tienen como objeto orientar en la administración, resguardo y control de calidad de los datos recabados; de entre los diez principios de vigilancia del clima destaca lo relativo a la gestión de la calidad y homogeneidad de los datos que deben evaluarse periódicamente y asimismo sobre el empleo de los sistemas de gestión de datos que facilitan la consulta, el uso y la interpretación de datos y productos que constituyen los elementos esenciales de los sistemas de vigilancia del clima [2].

La guía OMM-N° 100 define un CDMS como un sistema informatizado integrado que facilita el archivo, gestión, análisis, suministro y aprovechamiento eficaces de una amplia gama de datos climáticos integrados. Un CDMS está conformado [4] por los siguientes componentes (Fig. 1):

1. Series temporales de datos climáticos; se refiere a la capacidad de gestionar las variables relevantes con la observación del clima, por ejemplo, para la atmósfera son: temperatura y humedad del aire, velocidad y dirección del viento, etc.; incluye también los metadatos climáticos que describen las condiciones en que se dieron las observaciones y cómo se han gestionado los datos.
2. Gobernanza de los CDMS; son un conjunto coherente de políticas y procesos de gestión que aportan una base sólida para el establecimiento y el tratamiento de fuentes autorizadas de datos climáticos y servicios conexos.
3. Gestión de datos; actividades de introducción, extracción y rescate de datos por medio de la digitalización o la captura directa; el control de calidad define los métodos a utilizar como la verificación heurística, estadística o espacial; respecto al

Fig. 1. Componentes principales un CDMS [2].

aseguramiento de la calidad, indica los procesos, algoritmos y los mecanismos usados.

4. Suministro de datos; funcionalidad requerida para la entrega de datos climáticos en los formatos solicitados; incluye además los conceptos de descubrimiento de datos y metadatos climáticos.
5. Análisis de datos; procesos utilizados para el análisis de los datos, los algoritmos usados para la generación de datos derivados de las observaciones, por ejemplo, imágenes ráster para representar la distribución espacial de la temperatura; también se indicarán los mecanismos de procesamiento y análisis de datos y metadatos necesarios para desarrollar series temporales homogéneas de alta calidad.
6. Presentación de los datos; funcionalidades para visualizar los datos climáticos, puede referirse a la tecnología, SW o procesos empleados; por ejemplo, presentación en formatos tabulares, rosa de los vientos, fotografías o imágenes; la funcionalidad de descarga de datos provee un mecanismo de entrega de la información a los interesados.
7. Infraestructura de tecnología de la información, representa las funcionalidades necesarias para soportar y mantener un CDMS; definiéndose en este componente los sistemas de gestión de bases de datos adecuados.

3.1. Control de calidad de los datos

El control de calidad (CC) consiste en verificar si los valores obtenidos son representativos de la medición, los datos no se considerarán aptos para su uso o almacenamiento hasta que no hayan sido validados con cierto nivel de calidad [2] y precisamente un CDMS facilita estas tareas.

Los principios generales de calidad plantean principalmente que estos procesos que pueden realizarse de forma manual, automatizarse total o parcialmente, deben abarcar todo el ciclo de vida de los datos climáticos (desde la instalación de la EMA hasta el archivo final) conservando además las observaciones originales, recabando información de las fuentes de error y documentando exhaustivamente los métodos y prácticas empleadas.

Los pasos para el aseguramiento de la calidad [5] se describen a continuación:

1. Observación. Incluye los aspectos relativos al emplazamiento, instalación, captura de los datos, operación y mantenimiento de la EMA y los metadatos relacionados;

también se recomienda registrar los errores detectados y las medidas tomadas para resolverlos.

2. Entrega y captación de datos. Comprende la supervisión de la transmisión de los datos desde la EMA hasta su recepción en el CDMS.
3. Gestión de la BD climática. Se ocupa de la aplicación diferida del CC mediante las verificaciones manuales, automáticas o semiautomáticas, se recomienda el uso indicadores de calidad para informar si los datos han sido procesados, si se considera valioso, sospechoso, dudoso, erróneo, si ha sido modificado o estimado.
4. Archivo final. Lo relativo al resguardo final de la BD del CDMS, los datos archivados se acompañarán del registro de auditoría, el registro del CC aplicado y la documentación asociada.
5. Recuperación en caso de desastre. Describe lo referente al uso de copias de seguridad y almacenamiento adecuado para su aprovechamiento a perpetuidad.

3.2. Homogeneización de los datos

El objetivo es que los datos sean homogéneos, esto es, que sean uniformes o de "la misma naturaleza". Por causas ajenas al clima, la mayoría de los datos climáticos no son homogéneos [6], por ello se recomienda su verificación antes de realizar cualquier cálculo o generar cualquier producto climático.

Están disponibles dos métodos principales de homogeneización, 1) el método estadístico es usado cuando el comportamiento de una estación es claramente distinto a estaciones vecinas y 2) el método físico, cuando los ajustes se estiman mediante una relación física entre variables distintas, aunque requiere de datos más exactos y en ocasiones de observaciones con instrumentos que no se tienen en todos los emplazamientos.

4. El proceso de Software personal o PSP

En 1991 el Instituto de Ingeniería de Software de la Universidad Carnegie Mellon publica la primera versión del Capability Maturity Model for Software (Software CMM) que es un marco de madurez de procesos de SW orientado a las organizaciones [7]. Después de dirigir el desarrollo inicial de CMM, Watts S. Humphrey utiliza esos mismos principios en su propio trabajo en programas de tamaño de un módulo aplicando CMM hasta el nivel 5.

En 1994 Humphrey publica un manuscrito orientado a la enseñanza de estudiantes de ingeniería sobre el Proceso Personal de Software o PSP (Personal Software Process); este se enfoca en las prácticas de trabajo de los ingenieros individuales para ayudarlos a utilizar prácticas de ingeniería sólidas [8].

4.1. Definición de PSP

De acuerdo con Humphrey [9], el PSP, es un proceso de auto mejora que ayuda a controlar, administrar, y mejorar la forma de trabajar. Es un marco estructurado de formularios, directrices y procedimientos para el desarrollo de software.

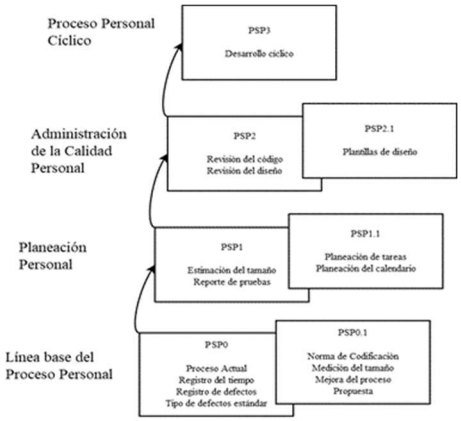


Fig. 2. Evolución de PSP, adaptado de [11].

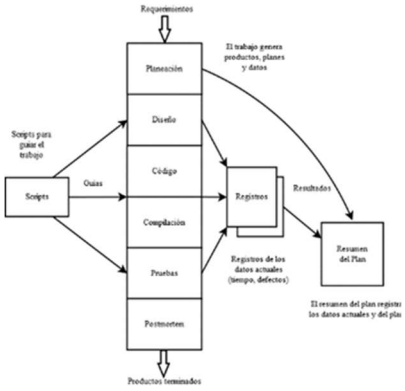


Fig. 3. Flujo del proceso PSP, adaptado de [11].

La estrategia del PSP consiste en motivar a cada ingeniero que adopte métodos eficaces; puesto que cada ingeniero es diferente, cada uno debe planificar su trabajo basándose en sus datos personales; utilizando procesos bien definidos y medidos, haciéndose responsables de la calidad de sus productos; además, es necesario detectar y corregir los defectos en etapas tempranas.

La medición del tiempo invertido en los productos de SW ayuda a conocer su rendimiento y posteriormente, utilizando los datos recabados, mejorar sus procesos personales [8]. Un proceso personal proporciona a los individuos un marco para mejorar su labor y para realizar de forma consistente trabajo de alta calidad [10].

4.2. Las fases del proceso PSP

Un proceso de SW establece el marco técnico de gestión para aplicar métodos, herramientas y personas a las tareas de SW. La definición del proceso identifica roles, especifica tareas, establece medidas y provee criterios de entrada y salida para los pasos principales [5]. La estructura de un proceso PSP (Fig. 3), tiene tres fases [10]:

1. Planeación: produce el plan del trabajo, un formulario de planeación es la guía para elaborar y documentar el plan y, provee un formato coherente para los resultados.
2. Desarrollo: se realiza el trabajo mediante: a) la definición de requerimientos, b) diseño del programa, c) revisión del diseño y corregir defectos, d) codificación, e) revisión del código y corrección de errores, f) construir o compilar y corregir los errores y, g) pruebas y corrección de los errores.
3. Post mortem: se realiza una comparación del desempeño actual contra el plan, se registran los datos del proceso, se realiza un reporte y se documentan todas las ideas para mejorar los procesos.

PSP contiene cuatro elementos básicos: 1) scripts o secuencia de comandos que guían la ejecución de un proceso personal, 2) formularios para especificar y guardar los datos necesarios, 3) métricas para proveer información cuantitativa sobre el proceso y el producto analizándolos con la intención de mejorarlos y supervisar el cumplimiento de las metas y, 4) estándares o normas precisas para guiar el trabajo, recopilación y uso de datos para permitir la aplicación coherente de las métricas.

4.3. La Evolución del PSP

El PSP tiene un marco de madurez similar al de CMM (Fig. 2) y cada fase se describe en seguida:

PSP0 - Proceso de referencia. Proporciona un marco de referencia para trabajar con el primer programa y recopilar datos sobre el rendimiento del ingeniero, conocido como línea base y es necesaria para determinar el impacto del PSP en su trabajo. PSP0.1 es la mejora del PS0 al cuál se le añade un estándar de codificación, la medición de tamaño y la propuesta de mejora del proceso. El objetivo del PS0.1 es que se comprendan los principios de tamaño y se desarrollen los ejercicios 2 y 3 que usarán como herramientas de medición en el resto del PSP [11].

PSP1 - Planeación del proceso personal. El PSP1 añade la planeación; la estimación del tamaño y de los recursos además de un informe de pruebas. PSP1.1 agrega la planeación del calendario y el seguimiento del estado. Los objetivos del PSP1 y PSP1.1 son: que se aprecie la relación entre el tamaño de los programas y el tiempo empleado en su desarrollo, que se planteen metas alcanzables, contar con un plan y una manera de determinar su estado [11].

PSP2 – Gestión personal de la calidad. Su objetivo es mejorar la productividad para producir programas de alta calidad mostrando cómo usar los datos recabados de los defectos para reducirlos en los pasos de compilación y pruebas. PSP2 agrega las revisiones personales del diseño y del código que permiten encontrar los defectos antes de procesarlos. PSP2.1 aborda los criterios que debe contener el diseño al estar concluido e ilustra varias técnicas para su verificación; el mismo enfoque puede utilizarse en otros pasos como la especificación de requisitos, la documentación y las pruebas [11].

PSP3 - Proceso personal cíclico. Hasta el PSP2 se lleva a cabo la construcción de pequeños programas, para enfrentar el desarrollo de programas grandes se subdivide en programas más pequeños (del tamaño de PSP2) y se integran al programa más grande. PSP3 es un proceso cíclico que sigue los principios del modelo de espiral de Boehm,

cada iteración incluye diseño, código, compilación y pruebas (unitarias y de integración). PSP3 es adecuado para programas de hasta varios miles de líneas de código y su objetivo es introducir los principios del proceso de escalado de procesos; asegurando la calidad de cada ciclo de desarrollo es posible concentrarse en verificar el rendimiento del último incremento sin interferencia de defectos anteriores [11].

Estimación con PROBE, recolección de datos y calidad del producto. PROBE son las siglas de PROxy Based Estimating (estimación basada en proxies) que utiliza proxies u objetos para estimar el tamaño probable de un producto. Comienza por calcular los objetos necesarios para construir lo que se plasmó en el diseño; se basa en datos históricos de objetos similares de al menos tres programas anteriores y emplea regresión lineal para determinar el tamaño probable del producto final.

En PSP se registran los tiempos consumidos en cada fase del proceso y el tamaño de los productos expresado en líneas de código o LOC.

Al final del trabajo se realiza el análisis post mortem actualizando el Resumen del Plan del Proyecto con los datos reales, se calculan los datos de calidad y de rendimiento contrastando el desempeño actual con lo planeado y calendarizado.

Como parte de la evaluación de la calidad de los productos, se recolectan datos de los defectos inyectados en cada fase y se lleva un registro de la cantidad de defectos encontrados, solucionados y el tiempo invertido en solventarlos [8].

5. Solución propuesta

Esta problemática ya se ha abordado anteriormente desde distintos enfoques, por ejemplo, la Comisión Nacional del Agua en México desarrolló MCH [16] que por ser un sistema provisto por la OMM cubre las necesidades de gestión de datos climáticos de los SHMN miembros de la organización, pero no está disponible al público en general.

Por otro lado, Flores-Román presentó la guía para la implementación de sistemas de medición de contaminantes atmosféricos [17] en combinación con estaciones meteorológicas, sin abordar el desarrollo del SW resalta la importancia de la verificación y control de calidad de los datos y el resguardo de la información procesada.

Por su parte la herramienta hidro-informática de Salgado-Álvarez [18] detalla distintos métodos para el aseguramiento de la calidad de datos climáticos antes de su aprovechamiento, aunque no considera el almacenamiento de la información procesada. En cambio, Guerrero-Higeras propone en su Modelo M3 [19] el procesamiento de información de distintas fuentes considerando también el factor tiempo, permitiendo el seguimiento en tiempo real de fenómenos meteorológicos; no obstante, el control de la calidad de los datos no forma parte de los elementos que lo conforman.

Ahora bien, en este documento se formula el diseño de un CDMS (Fig. 4) que será el apoyo para la administración de la RAMM. Esta solución propone la construcción de un sistema web con una arquitectura de microservicios aprovechando sus ventajas como el desarrollo de sistemas distribuidos, la escabilidad y su enfoque para modularizar el sistema permitiendo, de ser necesario, el uso de distintos lenguajes de

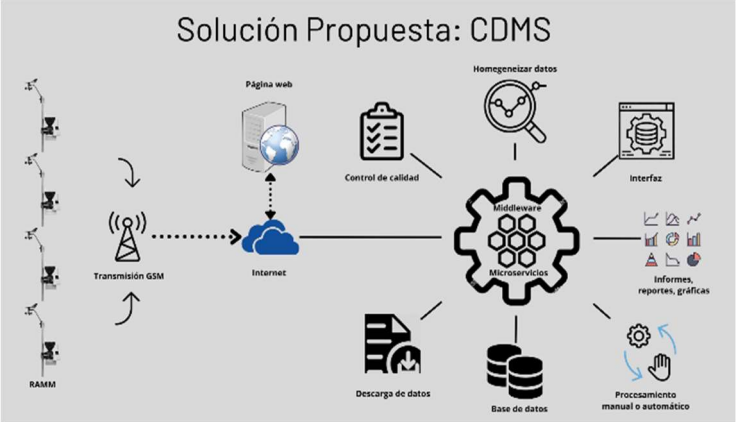


Fig. 4. Componentes de la propuesta de solución.

programación o tecnologías como sistemas middleware para realizar tareas recurrentes ya sea manuales o automáticas como la verificación de calidad o la homogeneización de los datos.

El CDMS deberá incluir los componentes señalados en [4] como *requeridos* excepto los obligatorios para miembros de la OMM ya que el DIAU no es miembro de esta organización.

Sobre el componente de Gobernanza al tratarse de normas o políticas de uso del sistema, tampoco se considera dentro del alcance de la realización del sistema; en cuanto al componente de infraestructura de TI, únicamente se considerará lo relativo a la gestión de la base de datos y su seguridad.

Bajo estas premisas, el CDMS respaldará las actividades de adquisición, validación y almacenamiento de la información recopilada y procesada, además, contará con herramientas para realizar cálculos climáticos y proveerá los mecanismos para la extracción, uso y aprovechamiento de los datos climáticos.

Asimismo, se recurre al marco de trabajo PSP que análogamente insta a dividir el desarrollo de un sistema grande en pequeños programas o módulos y al tiempo que se avanza en el desarrollo se recopilan los datos necesarios permitiendo al ingeniero de software potenciar su desempeño e incrementar la calidad de sus productos de SW.

6. Conclusiones

En este trabajo se han presentado las bases para la implementación y desarrollo de un Sistema de Gestión de Datos Climáticos para una red de estaciones meteorológicas automáticas según los estándares de la OMM.

Nuestra propuesta incluye la administración de las estaciones y los metadatos relacionados con la operación de la red, la verificación automática de calidad a los datos conforme se recibe de las estaciones y también a petición del usuario para los datos históricos, al menos una herramienta para la homogeneización de los datos y los mecanismos para abastecer al usuario con la información debidamente verificada por medio de una variedad de resúmenes y reportes gráficos y tabulares.

6.1. Trabajo futuro

El sistema descrito será realizado como parte del trabajo de maestría del primer autor, para lo cual se hará uso de Process Dashboard [12], una herramienta que facilita la adopción y el aprendizaje de PSP, que permite enfocarse en la realización del proyecto, en el análisis de productividad y calidad requeridos por PSP, incluye también los formularios, scripts [13] y de igual forma apoya en la realización de cálculo de las métricas arrojando los correspondientes reportes de productividad; un manual de la aplicación se puede consultar en línea en [14, 15].

En último término se deja la elección de la tecnología para el desarrollo del lado del servidor (backend) y del lado del cliente (frontend); en ambos casos deberán ser compatibles con la arquitectura propuesta. Para el backend, de entre las opciones está Node.Js o un conjunto de lenguajes actualmente usados: PHP, .Net, Python, Java o en su defecto un framework como Flask o SpringBoot. Para el frontend se cuenta con el lenguaje Javascript o un framework basado en este: Angular, Vue o inclusive la librería React.

Antes de comenzar con el desarrollo será necesario llevar a cabo la elicitación de requerimientos, su análisis y realizar un diseño general del CDMS; a partir de los cuáles se determinará el tamaño de cada módulo usando cada uno para escalar por las distintas fases de PSP. Al concluir el desarrollo del CDMS, se implementará y se hará una prueba de concepto para demostrar su utilidad con al menos un conjunto de datos de una estación de la RAMM.

Finalmente, se exhibirán los resultados obtenidos de Process Dashboard sobre el análisis de productividad para valorar las mejoras en el desempeño del desarrollador y concretar de que manera y cuanto evolucionó durante la ejecución del sistema descrito en el presente documento.

Referencias

1. INECC: ¿Qué es el clima? Instituto Nacional de Ecología y Cambio Climático (2018) www.gob.mx/inecc/acciones-y-programas/que-es-el-clima
2. Organización Meteorológica Mundial: Guía de prácticas climatológicas (OMM-N° 100), OMM, Ginebra (2018)
3. Barrell, S., Riishojgaard, L. P., Dibbe, J.: El sistema mundial de observación. Boletín, vol. 62, no. 1, pp. 9–16 (2013)
4. World Meteorological Organization: Climate data management system specifications. version 1.0, Ginebra, no. 1131 (2014)
5. Organización Meteorológica Mundial.: Directrices para el control de la calidad y el aseguramiento de la calidad de los datos de estaciones de observación en superficie para aplicaciones climáticas. Ginebra, no. 1269 (2021)
6. Organización Meteorológica Mundial: Directrices sobre homogeneización. Ginebra, n° 1245 (2022)
7. Paulk, M.: A history of the capability maturity model® for software. ASQ Software Quality Professional, vol. 12, no. 1 (2001)
8. Humphrey, W. S.: The personal software process (PSP). Software Engineering Institute (2018) doi: 10.1184/R1/6585197.V1
9. Humphrey, W. S.: PSP: Self-improvement process for software engineers. Addison-Wesley, New Jersey (2005)

10. Software Engineering Institute: The personal software process (PSP) body of knowledge. Version 2.0, Software Engineering Institute (2009)
11. Humphrey, W. S.: Introducing the personal software process. Annals of Software Engineering, vol. 1, no. 1, pp. 311–325 (1995) doi: 10.1184/R1/6585197.v1
12. Tuma Solutions: Process dashboard users manual (2022) www.processdash.com/sites/ processdash.com/ static/help/book.html
13. Software Engineering Institute: Personal software process (PSP) for engineers. Version 4, Course Materials (2018)
14. Tuma Solutions.: PSP Dashboard 2.6 (2022) www.processdash.com/
15. World Meteorological Organization (2022) community.wmo.int/mch-meteorology-climatology-and-hydrology-database-management-system
16. Flores-Román, D.: Guía de implementación de sistemas de medición de contaminantes atmosféricos. Tesis de maestría, Fundación Arturo Rosenblueth, Ciudad de México (2021)
17. Salgado-Álvarez, N.,: Diseño de una herramienta hidro-informática para el análisis de calidad de datos de estaciones meteorológicas automatizadas. Tesis de licenciatura, Universidad Autónoma del Estado de México, Estado de México (2018)
18. Guerrero-Higueras, Á.: Modelo de gestión de información meteorológica (M3) para calibración y validación de algoritmos, detección de riesgos meteorológicos y otras aplicaciones. Universidad de León, Castilla, España (2017)

ISSN 1870-4069

Implementación del módulo ESP32 como herramienta para el desarrollo de prácticas enfocadas al IoT

Ismael Minor Sampedro, Ricardo Álvarez González,
Rodrigo Lucio Maya Ramírez, Alba Maribel Sánchez Gálvez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Electrónica,
Laboratorio de Sistemas Digitales
México

ismael.minorsampedro01@gmail.com, {ricardo.alvarez,
rodrigo.maya, alba.sanchez}@correo.buap.mx

Resumen. Poco se habla de la aplicación del Internet de las Cosas en las instituciones educativas debido al crecimiento tan acelerado de esta tecnología, lo que en consecuencia presenta un rezago importante en la formación de sus estudiantes en estos tópicos, una forma de contribuir a la enseñanza de estos temas es mediante el modelo STEM donde los estudiantes sean capaces de adquirir conocimiento a través de la práctica, es por ello que en este trabajo se presenta el desarrollo de nodos dotados de sensores y actuadores que son capaces de procesar datos y comunicarse mediante Wi-Fi a través de la implementación del módulo ESP32, estos datos serán enviados mediante el protocolo MQTT a la plataforma Adafruit IO para poder registrar y desplegar de forma gráfica la información, todo con la finalidad de formar una Wireless Sensor and Actor Network que permita a los usuarios desarrollar prácticas y proyectos, monitorear variables, controlar procesos y proponer soluciones inteligentes de manera inalámbrica.

Palabras clave: IoT, protocolo MQTT, WSAN.

ESP32 Module Implementation as a Tool for Development of Practices Focused on IoT

Abstaract. Little is said about the application of the Internet of Things in educational institutions due to the rapid growth of this technology, which consequently presents a significant lag in the training of its students in these topics, a way of contributing to the teaching these topics is through the STEM model where students are able to acquire knowledge through practice, which is why this paper presents the development of nodes equipped with sensors and actuators that are capable of to process data and communicate via Wi-Fi through the implementation of the ESP32 module, these data will be sent via the MQTT protocol to the Adafruit IO platform in order to record and display the information graphically, all with the purpose of forming a Wireless Sensor and

Actor Network that allows users to develop practices and projects, monitor variables, control processes and propose intelligent solutions wirelessly.

Keywords: IoT, MQTT protocol, WSAN.

1. Introducción

1.1. El concepto de internet de las cosas

El concepto de Internet de las Cosas (IoT por sus siglas en inglés) fue introducido por primera vez en 1999 en una conferencia de Procter and Gamble por el ingeniero británico Kevin Ashton, donde hacía referencia a etiquetas de radiofrecuencia (RFID) que eran capaces de proveer información dentro de una cadena de suministros [1]. De este modo, al utilizar estas etiquetas, se reducía el riesgo de pérdida de información por la intervención humana.

Si bien el concepto de IoT ha estado presente por más de 20 años no ha sido hasta la última década que ha tomado importancia con el crecimiento exponencial de la tecnología y la llegada de la industria 4.0, en consecuencia, el concepto también ha evolucionado de modo que cada organización presenta una definición distinta, en la revista McKinsey, Chui, Loffer y Roberts, advierten que con la llegada del IoT hay un cambio en la forma en la que se genera información, es decir, en el mundo físico, los distintos elementos, objetos, o variables, con los que las personas interactúan de manera ordinaria, se convierten en proveedores o sistemas de información, esta definición considera la incrustación de sensores y actuadores en objetos físicos como, vehículos, edificios, marcapasos o wearables con capacidad de conexión a redes alámbricas o inalámbricas [2].

Es importante entender que estas definiciones no necesariamente difieren una de la otra, sino que al contrario, tienen un enfoque a los diferentes aspectos que engloba el fenómeno del IoT desde distintos puntos de vista y casos de aplicación, de este modo y para efectos de este trabajo, el concepto de IoT se entiende como la interconexión digital de objetos cotidianos a internet, es decir, los objetos ahora son capaces de generar información que pueda ser gestionada mediante medios computacionales en beneficio de los usuarios o procesos [3].

1.2. La demanda del IoT en México

De acuerdo con la Organización para la Cooperación y el Desarrollo Económicos (OCDE), mediante un ejercicio de exploración en 2016 de las direcciones IP de diferentes dispositivos, el volumen de IoT en México equivalía solamente apoco más de 8 millones de objetos conectados [4].

Sin duda, se trata de un área de oportunidad para la mejora en la calidad de vida y productividad de los individuos, empresas y gobiernos, así como a la mejor integración de estas tecnologías en nuestro día a día, sin embargo, su implementación no ha resultado una tarea sencilla.

Se considera que en México cerca de un 30 % de las compañías han comprendido las grandes ventajas que aporta el IoT, pues para estas es evidente que la integración

de una tecnología que permita generar nuevos modelos de negocio, obtener información en tiempo real de sistemas de misión crítica, diversificar las fuentes de ingresos, tener visibilidad global y mantener operaciones eficientes e inteligentes, es fundamental para poder ofrecer una ventaja competitiva [5].

Es claro que la demanda de la industria por las nuevas tecnologías va cada día en aumento, sin embargo, en México existe un gran desequilibrio entre las necesidades de la industria 4.0 y los programas educativos que incluyen tópicos como es el IoT, si bien, desarrollar un plan de estudios basado en este tema representa el camino a seguir, aplicar por esta alternativa implica una mayor cantidad de tiempo desde su propuesta hasta su implementación, tiempo que posiblemente la industria no puede esperar, es por ello que se ha optado por explorar nuevas metodologías como lo es el modelo STEM por sus siglas en inglés (Science Technology Enginering and Mathematics), este propone que los alumnos desarrollen aprendizaje basado en prácticas y proyectos donde se investigue y se diseñen soluciones a los problemas haciendo uso de tecnología moderna [6].

1.3. Sistemas embebidos disponibles en el mercado para la integración de proyectos

Actualmente en el mercado, una de las empresas lideres en el desarrollo de hardware embebido para la integración de proyectos es Mikroe, su objetivo es ayudar a desarrolladores a integrar soluciones tecnológicas, mediante la implementación de Clickboards™ y tarjetas de desarrollo [7].

A diferencia de lo existente en el mercado, este trabajo propone el desarrollo de módulos con sensores, actuadores o periféricos integrados, sin la necesidad de una tarjeta de desarrollo, ya que se implementará el SoC ESP32 en el propio módulo para poder procesar la información y al mismo tiempo ser capaz de enviar o recibir datos mediante la red Wi-Fi, adicionalmente, el módulo será capaz de operar con baterías, lo que favorece la implementación de estos dispositivos en distintas zonas de interés sin la necesidad de tenerlos concentrados en un solo lugar.

Además, teniendo como fundamento el modelo STEM, el objetivo principal radica en el desarrollo de hardware embebido que permita a los alumnos desarrollar prácticas de laboratorio con el fin de comprender integrar y proponer proyectos de IoT de una manera sencilla, modular y en el menor tiempo posible, haciendo uso de las plataformas más actuales en el mercado.

2. Desarrollo

Si la idea de conectar objetos entre sí y a internet no es nuevo, es razonable preguntar por qué el IoT es un tema que hoy en día está ganando popularidad.

El IoT tiene el potencial de cambiar fundamentalmente la forma en que interactuamos con nuestros alrededores. La capacidad de monitorear y administrar objetos en el mundo físico electrónicamente, hace posible llevar la toma de decisiones basada en datos a nuevos ámbitos de la actividad humana, a optimizar el rendimiento de los sistemas y procesos, ahorrar tiempo para las personas y las empresas, así como mejorar la calidad de vida [8].

Desde una perspectiva amplia, la confluencia de diferentes tendencias tecnológicas y de mercado [9] está permitiendo interconectar dispositivos más pequeños de forma económica y sencilla.

La tabla 1 muestra una clasificación de las herramientas que en la actualidad se encuentran disponibles para empezar a desarrollar aplicaciones relacionadas al IoT.

Analizando detenidamente la tabla 1 se observa que esta clasificación se realiza considerando los siguientes factores:

- Miniaturización.
  Los avances en la manufactura de circuitos integrados permiten desarrollar e integrar potentes sistemas como los SoCs (System on a Chip) y módulos en tamaños muy pequeños a gran escala y en consecuencia a bajo costo.
- Surgimiento de la computación en la nube.
  Cada día son más las plataformas que aprovechan recursos informáticos remotos conectados en red para procesar, gestionar y almacenar datos en la nube, de este modo es posible conectar dispositivos pequeños y distribuidos con el fin de interactuar con potentes sistemas de soporte que permitan el análisis y control de los datos y sistemas.
- Conectividad ubicua.
  La conectividad generalizada, de bajo costo y alta velocidad, sobre todo a través de servicios y tecnología inalámbricos con y sin licencia, hace que casi todo sea “conectable” [10].

2.1. Selección de Hardware

Para hacer una selección correcta del hardware es importante tomar en cuenta que los sistemas embebidos se clasifican en 4 tipos basados en el rendimiento y los requisitos funcionales: En tiempo real, independientes, en red y móviles.

En el desarrollo de este proyecto se plantea la creación de nodos Wi-Fi que faciliten la integración de proyectos enfocados al internet de las cosas a través de dispositivos modulares interconectados de manera inalámbrica, considerando estas características la clasificación corresponde a un sistema embebido en red, ya que estos están formados por componentes como controladores y sensores que se conectan a una red alámbrica o inalámbrica para realizar las tareas asignadas y dar salida a los dispositivos conectados.

Es por ello que para la implementación de estos nodos se hace uso del SoC ESP32 en su presentación de módulo, el cual integra todos los elementos necesarios para poder disponer de conectividad Wi-Fi, así como Bluetooth, del mismo modo dispone de una memoria para poder almacenar el firmware con en que se pretende desarrollar código, así como de GPIOs para poder trabajar con señales digitales o analógicas o establecer conexión mediante protocolos como UART, I2C o SPI.

2.2. Firmware y Software

- MicroPython.
  Para realizar la programación de estos módulos se utiliza Micropython el cual es una implementación del lenguaje de programación Python 3, optimizado para poder ejecutarse en un microcontrolador.

Tabla 1. Hardware y software disponible para IoT.

	Módulos	**Tarjetas de desarrollo**	**Firmware, SDK e IDE**	**Plataformas en la nube**
				
-ESP32-S ESP32-S2 ESP32-S3 -ESP32-C ESP32-C3 -ESP32 -ESP8266 -ESP8285	-ESP32-S -ESP32-C -ESP32 -ESP8266	-ESP32-DevKit -ESP-EYE -ESP Audio DevKits -ESP32-GoogleCloud IoT Kit -Esp32-Azure IoT kit	-ESP IDF -ARDUINO IDE -MICROPYTHON -ECLIPSE -VISUAL STUDIO	-Adafruit IO -Ubidots -ThingSpeak -Arduino Iot Cloud -Node RED

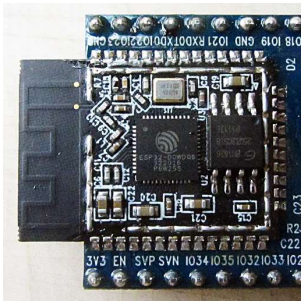


Fig. 1. Espressif ESP-WROOM-32 Módulo Wi-Fi & Bluetoot.

– Thonny.
Thonny es un programa muy interesante para empezar a aprender Python ya que engloba tres herramientas fundamentales: intérprete, editor y depurador.

2.3. Selección de la plataforma de servicio en la nube

Adafruit IO es un servicio en la nube que permite monitorizar datos y realizar paneles de control *online* también conocidos como *dashboards*, todo comunicado con el sencillo y eficaz protocolo MQTT.

2.4. Protocolo MQTT

Existen diferentes protocolos que permiten transmitir mensajes entre los dispositivos IoT en pequeños lapsos de tiempo; dentro de los más utilizados en este tipo de comunicaciones son *Message Queue Telemetry Transport* (MQTT), *Constrained Application Protocol* (CoAP), *Advanced Message Queuing Protocol* (AMQP) e *Hypertext Transfer Protocol* (HTTP) [11], algunos exigen mayor cantidad de recursos que otros o utilizan mayor ancho de banda.

Tabla 2. Datos registrados del Módulo Wi-Fi de salida (Relay).

value	feed_id	created_at	lat
1	2240365	2022-06-30 16:03:58 UTC	
1	2240365	2022-06-30 16:04:03 UTC	
1	2240365	2022-06-30 16:04:08 UTC	
1	2240365	2022-06-30 16:04:13 UTC	

Tabla 3. Datos registrados del Módulo Wi-Fi con sensor integrado (LDR).

value	feed_id	created_at	lat
2.881563	2238517	2022-06-30 16:03:58 UTC	
2.564102	2238517	2022-06-30 16:04:03 UTC	
2.515262	2238517	2022-06-30 16:04:08 UTC	
3.125763	2238517	2022-06-30 16:04:13 UTC	

Para seleccionar el protocolo de mensajería adecuado, es importante tener claro el objetivo del sistema IoT y sus requerimientos al momento de enviar mensajes o datos. Uno de los protocolos de comunicación más utilizados por dispositivos IoT es MQTT, ya que se trata de un protocolo de mensajería ligero para usar en casos de clientes con recursos limitados en cuanto al ancho de banda o consumo energético.

Se utiliza principalmente para comunicaciones de máquina a máquina (M2M) o conexiones del tipo IoT además de ser compatible con las plataformas disponibles como Adafruit IO.

La implementación del protocolo MQTT requiere de tres elementos fundamentales, los clientes, los *topics* y el *broker*.

Para el caso de este trabajo los nodos que integran el ESP32 toman el papel de clientes, la plataforma de Adafruit IO es el *broker*, por último, el *topic* es el elemento a que el cliente se suscribe o publica la información ya sea de un sensor o de un actuador.

Una de las ventajas del protocolo MQTT es que es posible conectar tantos clientes como lo permita el *broker* y estos pueden publicar y suscribirse a distintos *topics* de manera asíncrona permitiendo monitorear variables o accionar elementos dentro de una misma red.

Dicho lo anterior esto da paso a un concepto muy interesante conocido por sus siglas como WSAN.

2.5. Wireless Sensor and Actor Network

Las redes inalámbricas de sensores y actuadores por sus siglas en inglés (WSAN) se refieren a un grupo de sensores y actuadores conectados por un medio inalámbrico para realizar tareas distribuidas de detección y actuación.

En las WSAN, los sensores recopilan información sobre el mundo físico, mientras que los actuadores toman decisiones y luego realizan acciones apropiadas sobre el medio ambiente, lo que permite al usuario monitorear y actuar de manera efectiva a distancia [12].

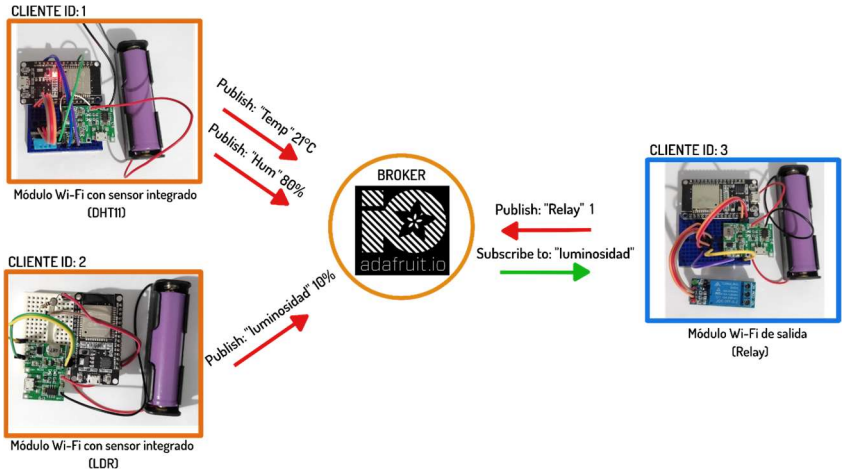


Fig. 2. Montaje de la WSAN con los módulos con sensor integrado y el módulo de salida.

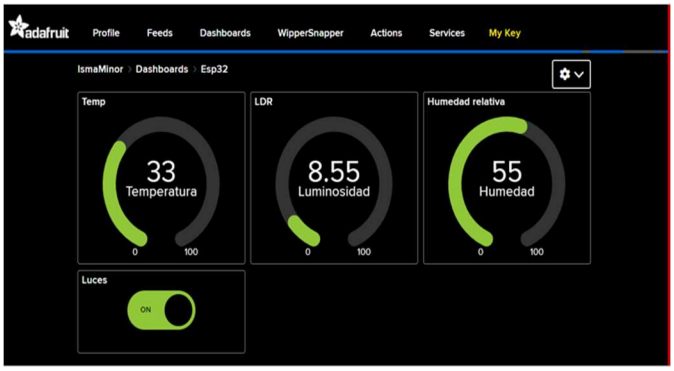


Fig. 3. Visualización de los datos en la plataforma de Adafruit IO.

3. Pruebas y resultados

Una vez analizados todos los elementos por separado el objetivo de este proyecto es desarrollar un kit de módulos Wi-Fi divididos en dos categorías: “Módulos Wi-Fi con sensor integrado"” y “Módulos Wi-Fi de salida”, estos se comunicarán mediante el protocolo MQTT para conformar así una *Wireless Sensor and Actor Network*, todo con la finalidad de facilitar la rápida integración de sistemas IoT de manera modular y sobre todo sin complejos sistemas de cableado.

3.1. Montaje de la WSAN

Para realizar las pruebas preliminares se hace uso de las tarjetas de desarrollo ESP32 *Devkit v1* por su practicidad para programar y realizar las conexiones, como se puede apreciar en la figura 2, se dispone de dos módulos con sensor integrado, uno de ellos

contiene un sensor DHT11 para el monitoreo de la temperatura y la humedad relativa, el otro módulo dispone de una fotorresistencia LDR que permite monitorear la intensidad de luz, en la parte derecha de la figura se aprecia el módulo Wi-Fi de salida que en caso de recibir alguna orden activa o desactiva un relevador, así también se muestra la estructura en cómo están distribuidos los módulos y la forma en que se comunican haciendo uso del protocolo MQTT, de este modo y para pruebas futuras si se desea integrar más módulos a la red solo basta con identificar el cliente, conectarse al *broker*, definir un *topic* y decidir si se va a publicar o suscribir información, esto último se configura mediante código.

3.2. Creación del Dashboard en Adafruit IO

Una vez programados los módulos la información que estos recolectan es enviada a la plataforma de Adafruit IO, no obstante, esta se envía de manera separada según el *feed* o *topic* a que el cliente publique o suscriba, para poder visualizar todo en un solo lugar Adafruit IO permite la creación de paneles también conocidos como *dashboards* los cuales reúnen toda la información de la WSAN y la muestran de una manera gráfica y estilizada.

3.3. Evaluación de los resultados

Las pruebas y resultados presentados a continuación, fueron evaluados dentro de un laboratorio de comunicaciones digitales donde se provee una red Wi-Fi de 2.4 GHz a la cual se conecta cada uno de los módulos, el objetivo de esta evaluación preliminar es comprobar si es posible integrar una WSAN teniendo como microcontrolador el módulo ESP32, de ser esto posible se verifica que el protocolo de comunicación MQTT funcione adecuadamente con la plataforma de Adafruit IO y finalmente se realiza un ejercicio de automatización mediante un "Módulo Wi-Fi con sensor integrado" y un "Módulo Wi-Fi de salida".

Uno de los principales retos al implementar una WSAN, es evitar la colisión de datos, para evitar esta situación al implementar el protocolo MQTT se debe identificar que, si bien se puede publicar o suscribir a un mismo *topic*, cada cliente debe tener un nombre o identificación única, de este modo es posible establecer esta comunicación y el *broker* no tendrá conflictos al momento de recibir la información de diferentes clientes.

Como se puede identificar en las Tablas 2 y 3, los datos registrados por el *broker* corresponden a dos dispositivos diferentes, no obstante, los datos fueron registrados en tiempos similares como se muestra en la columna "*created at"* respectivamente, esto

es posible ya que cada dispositivo o cliente tiene un ID único, esto evita que se genere una pila de datos o exista colisión de los mismos.

La recopilación de estos datos no solo comprueba que los módulos se comunican adecuadamente, sino que al trabajar con una plataforma en la nube esta información es almacenada y puede ser utilizada para evaluar los cambios ocurridos en el entorno donde se encuentran montados los módulos, tal como se muestra en la gráfica de la figura 4, donde se monitorean los cambios de temperatura dentro de una habitación por un lapso de 9 días, si se observa existe un comportamiento similar entre cada día que transcurre, si se toma un mayor número de muestras y se analizan los datos se abre una

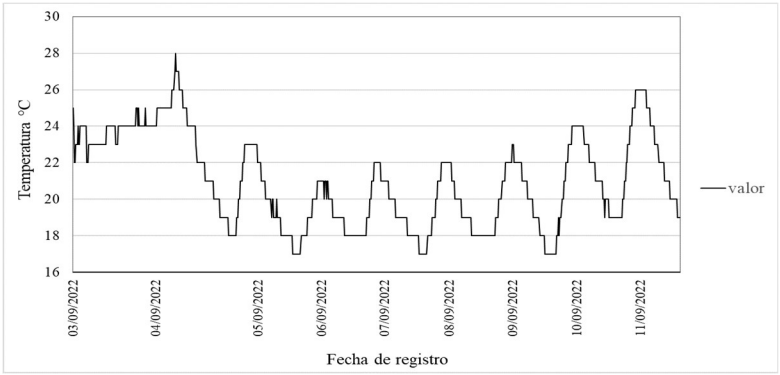


Fig. 4. Gráfica de los datos recopilados por el Módulo Wi-Fi con sensor integrado (DHT11).

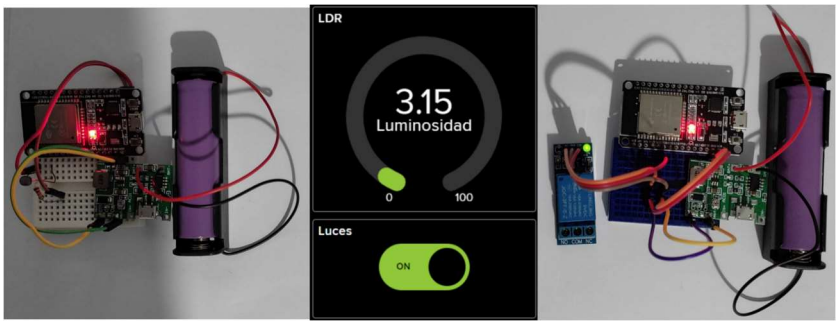


Fig. 5. Visualización de los datos en la plataforma de Adafruit IO y en los módulos cuando el relevador se acciona.

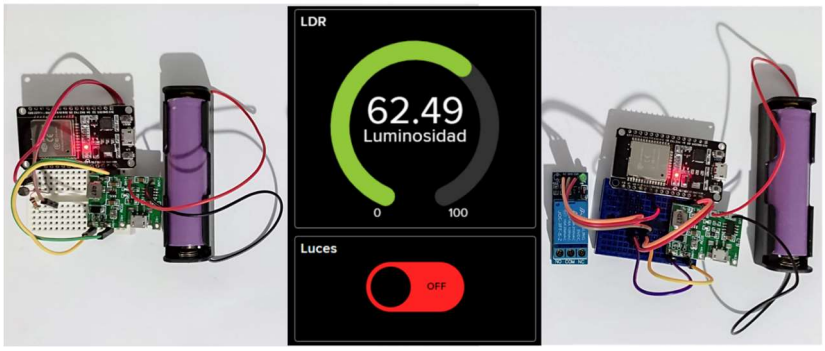


Fig. 6. Visualización de los datos en la plataforma de Adafruit IO y en los módulos cuando el relevador se desactiva.

oportunidad para desarrollar técnicas de mantenimiento predictivo para el sistema de ventilación de la habitación.

Otra de las pruebas realizadas es con la finalidad de facilitar la automatización de procesos de modo que la intervención humana sea mínima o en su defecto no sea necesaria.

Para este ejercicio se hace uso del Módulo Wi-Fi con sensor integrado (LDR) y del Módulo Wi-Fi de salida (Relay), en este caso el módulo con el LDR publica los datos al *topic* "luminosidad"' y el módulo con el relevador está suscrito a este mismo *topic*, de este modo es posible programar una condición:

Si el valor de "luminosidad" es menor a 10 acciona el relevador, en el caso contrario de que el valor sea mayor a 10 entonces se desactiva el relevador.

Las figuras 5 y 6 muestran el resultado de esta prueba.

4. Conclusiones y trabajo futuro

Durante el desarrollo de este trabajo ha sido posible identificar la importancia de desarrollar herramientas y métodos que permitan crear soluciones haciendo uso del Internet de las Cosas, afortunadamente hoy en día con las mejoras en temas de conectividad y la accesibilidad a potentes dispositivos electrónicos como lo es el módulo ESP32, se abre una ventana de oportunidad donde el concepto de IoT pasa de ser una idea a algo tangible, de modo que es posible conectar el mundo real con el digital mediante la implementación de redes dotadas de sensores y actuadores que se comunican en armonía mediante protocolos establecidos como lo es el MQTT.

La integración de estos módulos en la enseñanza aplicada bajo el modelo STEM, permite proponer un sistema asequible para que el usuario pueda adquirir aprendizaje a través del desarrollo de prácticas y al mismo tiempo sea capaz de proponer soluciones haciendo uso de este hardware.

Es importante mencionar que el proyecto se encuentra en una primera etapa, no obstante como trabajo futuro se diseñarán y fabricarán los PCB con el objetivo de tener nodos compactos que integren el módulo ESP32, un sensor o actuador específico, así como una etapa que permita suministrar la energía mediante baterías recargables, de este modo comparado con otros productos disponibles en el mercado estos dispositivos tienen la capacidad de poder montarse en diversas áreas de interés sin la necesidad de tener concentrado todo en una sola tarjeta de desarrollo.

Todo con la finalidad de implementar redes WSAN más grandes y con la facilidad de monitorear un mayor número de variables, así como tener la capacidad de accionar distintos dispositivos.

Referencias

1. Ashton, K.: That internet of things, thing: In the real world things matter more than ideas. RFID Journal Advances in Internet of Things, vol.6, no.4 (2016)
2. Chui, M.; Löffler, M., Roberts, R.: The internet of things. McKinsey Global Institute, pp. 12–13 (2010)
3. Román-Gallardo, A.: El internet de las cosas y su impacto en la educación. Colima, Universidad de Colima, pp. 7–8 (2020)
4. OECD: Internet access (indicator) (2022) doi.org/10.1787/69c2b997-en
5. Ortiz, G.: México rezagado en internet de las cosas. Deloitte México (2021) www2.deloitte.com/mx/es/pages/dnoticias/articles/internet-de-las-cosas-en-mexico.html
6. Florida Department of Education.: Defining STEM www.fldoe.org/academics/standards/subject-areas/math-science/stem/defining-stem.stml
7. Mitrovic, A.: Click Boards (2022) www.mikroe.com/click-boards

8. Manyika, J., Chui M.: The internet of things: mapping the value beyond the hype. McKinsey Global Institute, pp. 11 (2015)
9. Conant, S: The IoT will be as fundamental as the Internet itself. O'Reilly Radar (2015) radar.oreilly.com/2015/06/the-iot-will-be-as-fundamental-as-the-internet-itself.html
10. Poslad, S. Ubiquitous computing: smart devices, environments and interactions. Wiley (2009)
11. Serozhenko, M.: Mqtt vs. http: which one is the best for iot? (2017) medium.com/mqtt-buddy/mqtt-vs-http-which-one-is-the-best-for-iot-c868169b3105
12. Akyildiz, I. F., Kasimoglu, I. H.: Wireless sensor and actor networks: research challenges. Ad Hoc Networks, vol. 2, no. 4, pp. 351–367 (2004) doi: 10.1016/j.adhoc.2004.04.003

ISSN 1870-4069

Análisis del módulo de comunicaciones FiPy

Nicolás Quiroz-Hernández, José J. Medina-García,
Selene E. Maya-Rueda, Aideé Montiel-Martínez

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Electrónica,
México

{nicolas.quirozh, selene.maya, aidee.montiel}@correo.buap.mx,
jose.medinag@alumno.buap.mx

Resumen. El presente trabajo busca facilitar el uso e implementación del módulo de comunicaciones FiPy que cuenta con 5 tipos de tecnologías de comunicación diferentes, diseñado principalmente para IoT. El internet de las cosas (IoT) es una red colectiva de dispositivos conectados y a las tecnologías de comunicación a estos dispositivos entre sí y la nube. Las redes LPWAN permiten transmitir datos mediante comunicación inalámbrica entre dispositivos separados por kilómetros. Se realizó un análisis de algunos dispositivos que integran estas diferentes tecnologías de comunicación. Este dispositivo está dirigido a proyectos en los que la facilidad de intercambiar sistemas de comunicación se vital.

Palabras clave: Pycom, Redes IoT, LoRaWAN, Sigfox, LTE-M, WiFi, Bluetooth.

Analysis of the FiPy Communications Module

Abstract. Deaf community uses sign language as its main form of communication; however, most of the speaking community does not know how to understand that language, therefore the sign language recognition through technological developments has been an area of great interest for years. In this work, a proposal for this problem is presented, where regions of interest detection, manual and non-manual features extraction are carried out and for the recognition some BiLSTM networks with different architectures are used. The results obtained are an 73.99% accuracy, which are promising for the upcoming experiments. Finally, various actions are presented with the aim of improving the results as future work.

Keywords: Pycom, IoT networks, LoRaWAN, Sigfox, LTE-M, WiFi, Bluetooth.

1. Introducción

El IoT o internet de las cosas, se refiere a una red colectiva de dispositivos conectados y a las tecnologías que comunican a estos dispositivos entre sí y la nube [1]. En los últimos años, el internet de las cosas se ha convertido en una de las tecnologías

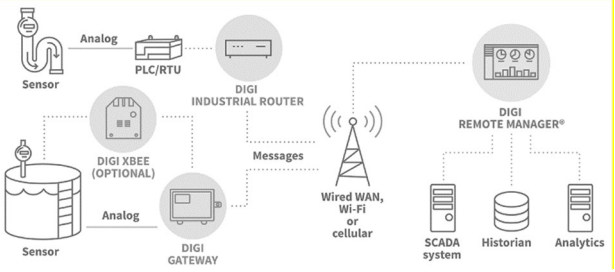


Fig. 1. Etapas de un sistema IoT [4].

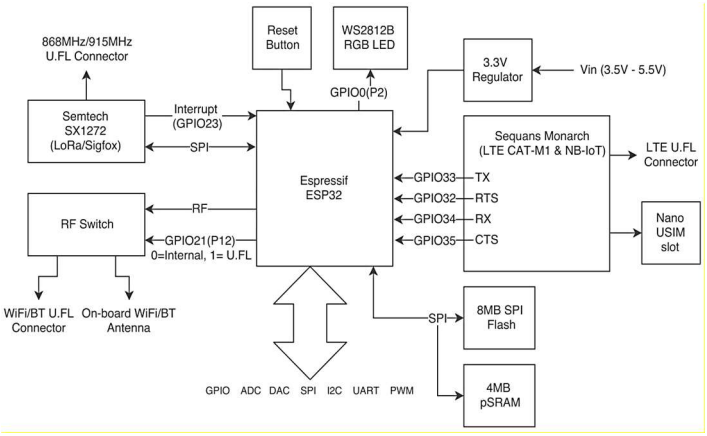


Fig. 2. Diagrama de bloques del módulo FiPy.

```
from network import WLAN
wlan = WLAN()

#Configuración de SSID y contraseña
wlan.init(ssid="FiPy", auth=(WLAN.WPA2, "Pycom_FCE"))
#id=1 indica que el modulo trabaja en modo WiFi AP
print(wlan.ifconfig(id=1))
```

Fig. 3. Código para crear un punto de acceso WiFi con SSID y contraseña propios.

más importantes en la actualidad, ya que podemos conectar objetos cotidianos como electrodomésticos, automóviles, termostatos, monitores para bebés, cámaras, etcétera, a internet a través de dispositivos integrados [2].

Internet se está desarrollando rápidamente y como resultado, Internet de las cosas se ha convertido en una realidad. La gloria de esta tecnología reside principalmente en todas las aplicaciones y oportunidades que ofrece para mejorar el día a día de las personas y los entornos empresariales. En la figura 1 se muestran los 5 componentes básicos de un sistema IoT, los cuales son sensores y actuadores, controladores, sistemas de comunicación, la nube y el monitoreo.

Por otro lado, para buscar dispositivos IoT se debe tener en consideración aspectos como bajo consumo energético y que sean pequeños, de ahí que los SoCs sean una parte importante de estos dispositivos [3].

```
from network import WLAN
import machine
wlan = WLAN(mode = WLAN.STA)

wlan.connect(ssid='Xiaomi', auth=(WLAN,WPA2,'123456'))
while not wlan.isconnected():
    machine.idle()
print("WiFi CConexión exitosa")
print(wlan.ifconfig())
```

Fig. 4. Código de conexión a red WiFi existente.

```
from network import Bluetooth
from machine import Timer

battery = 100
update = False
def conn_cb(chr):
    events = chr.events()
    if events & Bluetooth.CLIENT_CONNECTED:
        print('Cliente conectado')
    elif events & Bluetooth.CLIENT_DISCONNECTED:
        print('Cleinte desconectado')
        update = False
def chr1_handler(chr, data):
    global battery
    global update
    events = chr.events()
    print("eventos: ",events)
    if events & (Bluetooth.CHAR_READ_EVENT | Bluetooth.CHAR_SUBSCRIBE_EVENT):
        chr.value(battery)
        print("Transmitiendo :", battery)
        if (events & Bluetooth.CHAR_SUBSCRIBE_EVENT):
            update = True
bluetooth = Bluetooth()
bluetooth.set_advertisement(name='FiPy', manufacturer_data="Pycom", service_uuid=0xec00)
bluetooth.callback(trigger=Bluetooth.CLIENT_CONNECTED | Bluetooth.CLIENT_DISCONNECTED, handler=conn_cb)
bluetooth.advertise(True)
srv1 = bluetooth.service(uuid=0xec00, isprimary=True,nbr_chars=1)
chr1 = srv1.characteristic(uuid=0xec0e, value='leer_desde_aqui')
chr1.callback(trigger=(Bluetooth.CHAR_READ_EVENT | Bluetooth.CHAR_SUBSCRIBE_EVENT), handler=chr1_handler)
print('Iniciar servicio BLE')
def update_handler(update_alarm):
    global battery
    global update
    battery-=1
    if battery == 1:
        battery = 100
    if update:
        chr1.value(str(battery))
update_alarm = Timer.Alarm(update_handler, 1, periodic=True)
```

Fig. 5. Código para servidor BLE, que envía niveles de batería cada segundo.

Las redes LPWAN (Redes de Área Amplia de Baja Potencia) permiten transmitir datos mediante comunicación inalámbrica entre un dispositivo y una estación base o gateway separados por kilómetros con un bajo consumo de energía [5], existen 4 tipos de redes lideres con esta tecnología LoRaWAN, Sigfox, LET CAT-M1 y NB-IoT; LoRaWAN que es gratuito y es apoyado por la comunidad, lo que hace que esta red solo pueda implementarse si algún usuario a instalado un gateway de forma fija y con acceso libre, los dispositivos de comunicación de esta red son fáciles de obtener y de bajo costo rondando entre 10 y 25 dólares [6].

La red de Sigfox no es gratuita tiene un costo de 16 dólares anuales, pero no hace falta implementar un gateway ya que ellos proveen de esta infraestructura y existe una amplia cobertura de esta red, los dispositivos para poder acceder a la red de Sigfox son

```
from network import Bluetooth
from machine import Timer

def char_notify_callback(char, arg):
    char_value = (char.value())
    print("Nuevo valor: {}".format(char_value))
bt = Bluetooth()
print('Iniciando escaneo de servicios BLE')
bt.start_scan(-1)
adv = None
while(True):
    adv = bt.get_adv()
    if adv:
        try:
            if bt.resolve_adv_data(adv.data, Bluetooth.ADV_NAME_CMPL)=="FiPy":
                conn = bt.connect(adv.mac)
                print("Connected to FiPy")
                try:
                    services = conn.services()
                    for service in services:
                        chars = service.characteristics()
                        for char in chars:
                            c_uuid = char.uuid()
                            if c_uuid == 0xec0e:
                                if (char.properties() & Bluetooth.PROP_NOTIFY):
                                    char.callback(trigger=Bluetooth.CHAR_NOTIFY_EVENT, handler=char_notify_callback)
                                    print(c_uuid)
                                    break
                except:
                    continue
        except:
            continue
bt.stop_scan()
bt.disconnect_client()
```

Fig. 6. Código para cliente bluetooth.

de costos más elevados entre 40 y 200 dólares, sin embargo, la mayoría de estos dispositivos cuentan con una suscripción gratuita de un año a esta red [7].

Las redes LTE CAT-M1 y NB-IoT son redes dedicadas para dispositivos que requieren datos durante largos periodos de tiempo en lugares de difícil acceso, su costo es más elevado ya que funciona de la misma forma que una red LTE con una cuota para mantener la red activa [8].

Pycom es una empresa de origen europeo que crea dispositivos que contienen controladores y sistemas de comunicación, proporciona un entorno completo de dispositivo a nube y herramientas de administración segura y confiable para implementar IoT [9].

FiPy es un módulo programable en lenguaje MicroPython que integra WiFi, Bluetooth de baja energía, LoRa, Sigfox y LTE-M dual, que es de gran apoyo cuando se quieren integrar proyectos que tienen la necesidad de ser versátiles en la comunicación de datos a una base de datos [9].

2. Desarrollo

Para programar el módulo FiPy se utiliza Visual Studio Code, este puede tener la función de ser controlador y sistema de comunicación al mismo tiempo, ya que integra un SoC de Espressif ESP32, en la figura 2 se muestra un diagrama de bloques de lo que integra el módulo FiPy, este cuenta con dispositivos de comunicación LPWAN, BLE y WiFi, tiene conectores para antenas de los diferentes sistemas de comunicación y adicionalmente cuenta con pines dedicados para el usuario.

```
from network import LoRa
import socket
import ubinascii
import struct
#Inicializamos el modo LoRaWAN para México 915 MHz
lora = LoRa(mode=LoRa.LORAWAN, region=LoRa.US915)
#Parametros de autentificació ABPn
dev_addr = struct.unpack(">l", ubinascii.unhexlify('00000005'))[0]
nwk_swkey = ubinascii.unhexlify('2B7E151628AED2A6ABF7158809CF4F3C')
app_swkey = ubinascii.unhexlify('2B7E151628AED2A6ABF7158809CF4F3C')
for i in range(0,8):
    lora.remove_channel(i)
for i in range(16,65):
    lora.remove_channel(i)
for i in range(66,72):
    lora.remove_channel(i)
lora.join(activation=LoRa.ABP, auth=(dev_addr, nwk_swkey, app_swkey))
s = socket.socket(socket.AF_LORA, socket.SOCK_RAW)
s.setsockopt(socket.SOL_LORA, socket.SO_DR, 5)
s.setblocking(True)
#Algunos datos a enviar
s.send(bytes([0x01, 0x02, 0x03]))
s.setblocking(False)
#Espera de datos Downlink
data = s.recv(64)
print(data)
```

Fig. 7. Programa de comunicación LoRaWAN en modo ABP.

2.1. WiFi

Los módulos de desarrollo de Pycom tienen una antena integrada, por lo que no es necesario utilizar una antena externa, el módulo puede trabajar de dos formas como punto de acceso donde el dispositivo crea una red local o como cliente donde el dispositivo se conecta a una red existente. El consumo energético de WiFi en modo punto de acceso es de 126 mA mientras que en modo cliente es de 137 mA, estas mediciones son proporcionadas por el fabricante.

2.1.1. Punto de acceso WiFi

Para configurar un punto de acceso básico se debe utilizar el código mostrado en la figura 3, este establecerá de manera predeterminada la dirección IP 192.168.4.1.

2.1.2. Cliente WiFi

Para conectar el módulo a una red existente, se debe configurar el modo WLAN.STA, en la figura 4 se muestra el código de programa básico para conectarse a una red existente.

2.2. Bluetooth

El módulo FiPy cuenta con Bluetooth 4.2 de baja energía y el consumo de este durante su uso es de 121 mA, esta medición es proporcionada por el fabricante. Se pueden escanear los dispositivos BLE cercanos y funcionar como cliente o servidor bluetooth.

```
from network import LoRa
import socket
import time
import ubinascii
#Inicializar modo LoRaWAN para México 915 MHz
lora = LoRa(mode=LoRa.LORAWAN, region=LoRa.US915)
#Crear parametros de autentificación OTAA
#Estos datos deben cambiar de acuerdo al dispositivo FiPy
app_eui = ubinascii.unhexlify('ADA4DAE3AC12676B')
app_key = ubinascii.unhexlify('11B0282A189B75B0B4D2D8C7FA38548B')
for i in range(0,8):
    lora.remove_channel(i)
for i in range(16,65):
    lora.remove_channel(i)
for i in range(66,72):
    lora.remove_channel(i)
lora.join(activation=LoRa.OTAA, auth=(app_eui, app_key), timeout=0)
while not lora.has_joined():
    time.sleep(2.5)
    print('No conectado...')
print('Conectado')
s = socket.socket(socket.AF_LORA, socket.SOCK_RAW)
s.setsockopt(socket.SOL_LORA, socket.SO_DR, 5)
s.setblocking(True)
#Envio de datos
s.send(bytes([0x01, 0x02, 0x03]))
s.setblocking(False)
#Espera de datos Downlink
data = s.recv(64)
print(data)
```

Fig. 8. Programa de comunicación LoRaWAN en modo OTAA.

```
from network import Sigfox
import socket
#Iniciar Sigfox para México zona RCZ2
sigfox = Sigfox(mode=Sigfox.SIGFOX, rcz=Sigfox.RCZ2)
s = socket.socket(socket.AF_SIGFOX, socket.SOCK_RAW)
s.setblocking(True)
s.setsockopt(socket.SOL_SIGFOX, socket.SO_RX, False)
#Envio de datos
s.send(bytes([1, 2, 3, 4]))
#Espera de datod Downlink
r = s.recv(32)
print(ubinascii.hexlify(r))
```

Fig. 9. Programa para enviar datos y esperar mensaje de bajada de Sigfox.

2.2.1. Servidor BLE

En la figura 5 se muestra el código para poner el módulo como servidor BLE, este programa espera la conexión de un cliente BLE para que pueda enviar el nivel de batería cada segundo disminuyendo constantemente.

Tabla 1. Comparación entre dispositivos que integran diferentes sistemas de comunicación.

Tecnología de comunicación	Dispositivo				
	FiPy	TTGO LoRa32	SIM7600	MKRFOX1200	SFM11R2D
Sigfox	x			x	x
LoRaWAN	x	x			
LTE			x		
LTE-M	x				
BLE	x	x			
WiFi	x	x			
Procesador	ESP32	ESP32		SAMD21 Cortex-M0	
Costo (MXN)	$1,985	$1500	$1900	$1300	$890

2.2.2. Cliente BLE

En la figura 6 se muestra un ejemplo de cliente bluetooth, que busca un servidor bluetooth llamado FiPy y se conecta a este, para obtener los datos del ejemplo anterior e imprimirlos en la consola de VS Code.

2.3. LoRa

Para poder conectarse a la red LoRaWAN es necesario contar con una puerta de enlace cercana a el nodo final en este caso al módulo FiPy, la gran ventaja de esta red es la cantidad de datos por mensaje que se pueden enviar.

Para utilizar el protocolo LoRa es necesario conectar la antena adecuada en el conector Ipex Ufl, existen dos formas básicas de acceder a la red de LoRaWAN ABP y OTAA.

En el modo ABP las claves de cifrado se configuran manualmente en el dispositivo y puede enviar tramas sin necesidad de un intercambio de claves, en la figura 7 se muestra un ejemplo de programa para poder realizar el envío de datos utilizando este método.

En el modo OTAA el módulo envía una solicitud de unión a la puerta de enlace LoRaWAN utilizando app_eui y app_key, si estas claves son correctas la puerta de enlace aceptara la unión, en la figura 8 se muestra un ejemplo de implementación de este modo.

2.4. Sigfox

La ventaja de la red Sigfox es su cobertura ya que no es necesario implementar una puerta de enlace, además de que el dispositivo FiPy cuenta con 1 año de suscripción a esta red, la cantidad máxima de datos que pueden ser enviados son 12 bytes por mensaje con un máximo de 140 mensajes cada 24 horas.

El módulo se puede configurar para funcionar en diferentes países en función de las zonas RCZ específicas, en la figura 9 se muestra un ejemplo de programa para poder enviar y esperar datos de bajada a través de Sigfox.

Tabla 2. Comparación entre dispositivos que integran el sistema de comunicación Sigfox.

Especificaciones	Dispositivo		
	FiPy	MKRFOX1200	SFM11R2D
Máximo poder Tx	+20dBm	+14.5dBm	+22.5dBm
Rango al nodo	>50 Km	>20 Km	>50 Km
Consumo	192 mA	130mA	170mA
Frecuencia	RCZ1-4	RCZ1	RCZ1-2

Tabla 3. Comparación entre dispositivos que integran el sistema de comunicación LoRaWAN.

Especificaciones	Dispositivo	
	FiPy	TTGO LoRa32
Máximo poder Tx	57 dBm	20 dBm
Rango al nodo	>40 Km	<30 Km
Consumo	156 mA	130 mA
Frecuencia	860-1020 MHz	868-915 MHz

Tabla 4. Comparación entre dispositivos que integran el sistema de comunicación LTE y LTE- M.

Especificaciones	Dispositivo	
	FiPy	SIM7600
Velocidad de datos	40-375 kbps	50 Mbps
Versión de LTE	CAT-M1, NB-IoT	CAT4
Consumo	420mA	530 mA
Frecuencia	699-2690 MHz	824-1910 MHz

Tabla 5. Comparación entre dispositivos que integran el sistema de comunicación BLE.

Especificaciones	Dispositivo	
	FiPy	TTGO LoRa32
Versión BLE	4.2	4.2
Máximo poder de Tx	+12 dBm	+9 dBm
Consumo	121 mA	130 mA
Frecuencia	30 MHz-12.5 GHz	30 MHz-12.5 GHz

Tabla 6. Comparación entre dispositivos que integran el sistema de comunicación WiFi.

Especificaciones	Dispositivos	
	FiPy	TTGO LoRa32
Estándar WiFi	802.11 b/g/n/e/i	802.11 b/g/n
Protección de acceso	WPA/WPA2/WPA2-Enterprise/WPS	WPA/WPA2/WPS
Frecuencia	2412-2484 MHz	2412-2048 MHz
Máximo poder de Tx	15 dBm	14 dBm
Consumo	137 mA	240 mA

2.5. LTE-M

FiPy es compatible con LTE CAT-M1 y NB-IoT, estos son los protocolos celulares más nuevos, de baja potencia y largo alcance. Al momento de escribir este trabajo estas conectividades no están ampliamente disponibles, sin embargo, el dispositivo tiene la posibilidad de poder implementarla ya que cuenta con el hardware para esta red.

3. Resultados

Para poder implementar el programa de LoRaWAN fue necesario crear un gateway con un SoC ESP32, ya que en la zona no existía alguno, y como ya se mencionó para las redes LTE-M, en México apenas se empieza con estas redes por lo que no existe cobertura tan amplia de estas.

En la tabla 1 se muestran algunos dispositivos que integran 1 o más de estas redes, comparando diferentes módulos existentes para las 5 redes integradas en este dispositivo.

En las tablas 2 a 6 se muestra una comparación de los diferentes dispositivos enlistados en la tabla 1, clasificados de acuerdo con la tecnología de comunicación que integran.

4. Conclusión

La implementación de este dispositivo es más eficiente y barato a la hora de integrar varias redes en un solo proyecto y resulta en un consumo energético más bajo que integrar varios módulos para diferentes redes. este módulo está dirigido a proyectos en los que la versatilidad de la comunicación es un factor importante, ya que puede cambiar de red en cualquier instante, sin embargo, si el proyecto solo requiere de una red para envío de datos o control basta con adquirir cualquiera de los otros dispositivos enlistados anteriormente.

Referencias

1. AWS Amazon: ¿Qué es IoT? (2022) aws.amazon.com/es/what-is/iot
2. ORACLE: ¿Qué es el IoT? (2022) www.oracle.com/mx/internet-of-things/what-is-iot
3. Gracia, M: Deloitte, IoT - Internet of Things (2022) www2.deloitte.com/es/es/pages /technology/articles/IoT-internet-of-things.html
4. DIGI (2022) es.digi.com/blog/post/the-4-stages-of-iot-architecture
5. Becolve Digital: LPWAN: qué son y para qué se utilizan (2022) www.m2mlogitek.com/lpwan-que-son-y-para-que-se-utilizan
6. LoRaWAN (2022) lorawan.es
7. Sigfox OG Technology: Sigfox oG technology by unabiz (2022) www.sigfox.com/en
8. NC Tech. The New TechCompany: ¿Qué es IoT de banda estrecha (NB-IoT)? (2022) nctech.com.mx/blog/iot-industrial/nb-iot/
9. pycom: Next Generation Internet of Things Platform (2022) pycom.io

ISSN 1870-4069

Vehicular Ad-Hoc Network Throughput Evaluation in 3D Environments

Josefina Castañeda-Camacho, Alejandro Sánchez-Mendoza,
Ana María Rodríguez-Domínguez, José Fermi Guerrero-Castellanos

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Electrónica,
Mexico

alejandro.sanchezme@alumno.buap.mx

Abstract. The deployment of the 5th generation of mobile communication networks (5G) brought countless opportunities to improve and create technologies for everyday use. Autonomous aerial vehicles such as drones (UAVs) or their terrestrial counterpart, rovers (UGVs) are systems that have appeared in both private and public sectors to offer different services. Vanets are self-organizing and dynamic networks designed specifically for vehicle communication in changing environments. This paper shows an analysis of one of the most important metrics in network design, the throughput, in a vanet network environment using Montecarlo discrete event simulations. The performance of different modulation methods and how the conditions of the communication channel affect the throughput are analyzed with the main goal to define which of the modulation methods is better for each channel conditions.

Keywords: Vehicular Ad-hoc networks, throughput evaluation, 5G.

1 Introduction

Since the appearance of the human species, communication processes have been essential for its survival and growth, in modern times this has not changed. An example of these are mobile communication networks that have been transformed since their invention in search of improving their performance and efficiency [1].

The fifth generation of mobile communication networks or 5G as it is commonly known, is today a reality and brings with it improvements in network speed, in the use of bandwidth and in the number of devices supported by the network, among others [9].

The use of both terrestrial (AGV) and aerial (UAV) autonomous vehicles to carry out different tasks in public and private sectors has increased in the last decade, this is mainly attributed to the fact that the price of these devices have become cheaper.

As a result of the growth of these technologies, new needs have arisen, one of which is to intercommunicate these vehicles efficiently to improve their performance and avoid any type of failure. Ad-Hoc vehicular networks (vanet) emerge as a solution to this need thanks to the benefits offered by 5G [2].

Table 1. Comparison of Parameters Considered by Referenced Works.

Reference	Parameters
[3]	Throughput / SIR
[7]	Throughput / SIR
[9]	Outage Probability

Vanets are a particular case of mobile Ad-Hoc networks (manets) focused on vehicular environments. A vanet is a wireless network characterized by allowing the nodes to communicate cooperatively, to exchange relevant information such as road conditions or situations that arise during a journey, providing a self-organized network environment that does not require an established infrastructure or centralized administration [7].

Ideally, a communications network should be simple, flexible, and cost-effective, but robust enough to support the traffic that flows through it. To find the conditions in which a network approaches optimal performance, it is necessary to carry out a process of network sizing.

The sizing of a network is the design process through which the minimum capacities that each segment of the network must have to satisfy the operation requirements and ensure the quality of service to users are determined [3, 8].

There are many parameters that can be measured during network sizing, one of the most important is the throughput, this metric gives an idea of the total quantity of date would be possible to transmit along the communication network [4]. Table 1 present a resume of the parameters used for different authors to evaluate vehicular adhoc networks.

This work is focused on the measurement of the throughput of a vanet. Present work is structured as follows. Second section presents the mathematical background used to develop the simulation scenarios. Third section includes the simulation set-up and results. Lastly conclusions are presented.

2 Mathematical Background

The simulation algorithm is composed by three fundamental calculations, first the power received by each mobile user inside the network working area and in the interference cells, secondly, the signal-to-interference ratio using the calculated power from the interfering users and lastly the throughput for each communication link according to its obtained SIR value.

Equation 1 shows the proposed propagation model, it is a modification of the free space propagation model, in this case the received power evaluation considers the shadowing effects, distance losses and transmitter antenna gain (as a function of the user position), this helps to measure the power losses related to the transmission of the data through the propagation medium:

$$P_{rx} = P_{tx} G_{tx} G_{rx} \frac{h_{tx} h_{rx}}{d^{\mu}} 10^{\frac{\lambda}{10}}, \quad (1)$$

where P_{tx} is the transmission power, G_{tx} and G_{rx} are the antenna gains for the transmitter and the receiver respectively, d is the distance between the antenna and

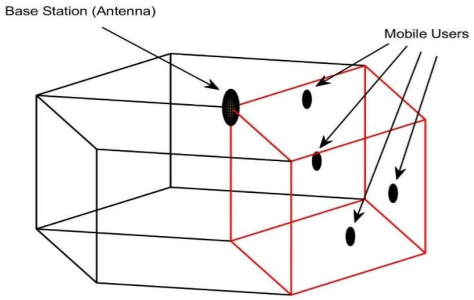


Fig. 1. Network Communication Cell.

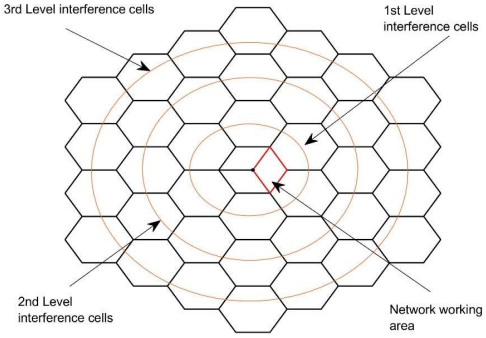


Fig. 2. Hexagonal Cell Regions.

the end device, μ is the propagation losses exponent, h_{tx} and h_{rx} are the transmitter and receiver antenna height , finally, λ is the characteristic Gaussian random variable of the log-normal distribution which models the shadowing effects.

The transmitted antenna gain is considered as a parabolic function which depends on the angle $\phi_{l,n}$ between the end device and the maximum transmission ray of the antenna, this function is expressed in equation 2:

$$G_{tx}(\phi_{l,n}) = \begin{cases} 1 - \frac{1-q}{\left(\frac{\pi}{3}\right)^2}\phi_{l,n}^2 & si|\phi_{l,n}| \leq \sqrt{\frac{1-p}{1-q}\frac{\pi}{3}} \\ p & si|\phi_{l,n}| > \sqrt{\frac{1-p}{1-q}\frac{\pi}{3}} \end{cases}, \quad (2)$$

where q represents the gain level of the antenna inside the maximum transmission sector and p is the mean normalized gain of the side lobes [5, 6].

Then, once the received power is calculated, the signal-to-interference is obtained as a quotient between the received power by the mobile user in the working area and the summation of the received power by the interfering nodes in the adjacent cells, as shown in equation 3:

$$SIR_i = \frac{P_{rx_i}}{\sum_{k=1}^{N_{int}} P_{rx_k}}, \quad (3)$$

Table 2. Transmission Parameters.

Parameter	Value
Transmission Power	89.1 mW
Sensitivity	1.9 pW
Bandwidth	10 MHz

where P_{rx_i} is the received power by $i-th$ mobile user inside the interest area, N is the number of interfering users and P_{rx_k} is the power received by the $k-th$ interfering user. Signal-to-interference ratio helps to understand the impact of the interference caused by nearly communications happening around the network working area.

Lastly, to calculate the communication link throughput, it is necessary to start from the link quality metric γ, given as a function of the bit energy and the spectral density of the interference, as shown in equation 4:

$$\gamma = \left(\frac{E_b}{I_0}\right)_i = \frac{\omega_0}{R_i} SIR_i, \tag{4}$$

where E_b is the bit energy, I_0 is the spectral density of the interference, ω_0 is the system bandwidth, R_i is the $i-th$ user throughput and SIR_i is the $i-th$ user Signal-to-Interference Ratio.

Throughput refers to the total quantity of bits per second transmitted form the sender to the receiver and it can be modified depending on the modulation scheme used by the system. Equation 5 shows how throughput is calculated:

$$R_i = \begin{cases} \left(\frac{\omega_0}{\gamma}\right) SIR_i & siSIR_i > m\gamma \\ m\omega_0 & siSIR_i \leq m\gamma \end{cases}, \tag{5}$$

where m is a factor related to the modulation scheme and γ is the channel quality metric, when γ increase it means the channel conditions are worst, communications standards implemented for vanets commonly permit four modulations, BPSK, QPSK, 16QAM and 64QAM; m takes 1, 2 ,4 and 6 as its value respectively for each modulation.

3 Simulation Set-Up

We have developed a discrete event Montecarlo simulation to evaluate the equation (5). For the simulation scenario, we are considering a communication environment divided in micro-cellular cells modeled as hexagonal prisms as shown in figure 1.

The network is formed by four mobile users' ash shown in figure 1. The simulation occurs in a single cell with a 10mts radius and 3 meters height. Also, considering a 120 degrees sectorization.

This sector matches with the antenna propagation pattern placed in the center of the cell. Similar scenarios are created for the adjacent cells and interfering users, this shown in figure 2. Results obtained from the simulation are shown in the following section.

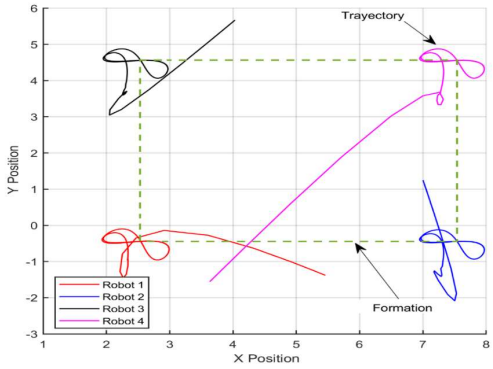


Fig. 3. Nodes Movement Trajectories.

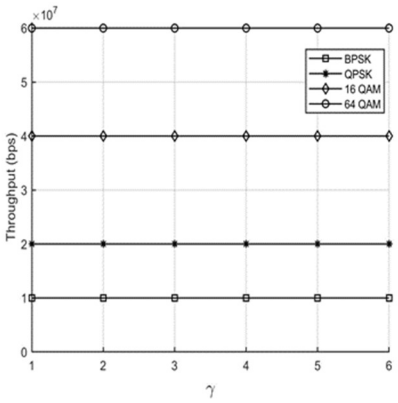


Fig. 4. Throughput Behavior for Node 1.

4 Results

4.1 Mobility

To simulate node mobility a control algorithm for trajectory and formation following was implemented, with random initial conditions, the nodes are commanded to bring a square formation and follow the Bernoulli's lemniscate trajectory, figure 3 shows the trajectory followed by the nodes in one of the events.

4.2 Throughput

Considering that the maximum transmission ray of the transmitting antenna matches with the X axe in 0 value, robot number 1 and 3 have a better positioning compared with robots 2 and 4, figure 4 shows the throughput behavior for robot 1.

In this case the bandwidth usage is the maximum possible for all four modulations and all channel conditions (γ), one of the reasons is that the distance between the user

Table 3. Throughput for $\gamma = 1$.

USER	BPSK	QPSK	16QAM	64QAM
1	10 Mbps	20 Mbps	40 Mbps	60 Mbps
2	10 Mbps	20 Mbps	40 Mbps	60 Mbps
3	10 Mbps	20 Mbps	40 Mbps	60 Mbps
4	10 Mbps	20 Mbps	40 Mbps	60 Mbps

Table 4. Throughput for $\gamma = 3$.

USER	BPSK	QPSK	16QAM	64QAM
1	10 Mbps	20 Mbps	40 Mbps	60 Mbps
2	10 Mbps	20 Mbps	39.6 Mbps	54 Mbps
3	10 Mbps	20 Mbps	40 Mbps	59.6 Mbps
4	10 Mbps	19.6 Mbps	28.9 Mbps	32.8 Mbps

Table 5. Throughput for $\gamma = 6$.

USER	BPSK	QPSK	16QAM	64QAM
1	10 Mbps	20 Mbps	40 Mbps	60 Mbps
2	10 Mbps	19.9 Mbps	34.6 Mbps	39.6 Mbps
3	10 Mbps	20 Mbps	39.6 Mbps	55.3 Mbps
4	9.9 Mbps	16.3 Mbps	20.6 Mbps	23.1 Mbps

and the antenna is short and losses are minimum, in the other hand, in figure 5, the throughput behavior for node 4 is shown.

The impact of the distance and the angle between the transmitter and the receiver is clear. Impacts increase when the channel conditions are poor and the modulation complexity increase, this let us know that modulation schemes such as 16 QAM and 64 QAM present a higher sensitivity to channel conditions than BPSK and QPSK, however, channel usage is better in 16 QAM and 64 QAM.

Finally, tables 2, 3 and 4 show throughput values for all modulations with channel conditions equal to 1, 3 and 6.

Once results are obtained it is important to compared them with results reported for similar works in the literature related. Table 6 shows a comparison between the results already presented and some relevant data detailed by other authors.

5 Conclusion

Results show that the throughput can be significantly affected by channel conditions, users' mobility, and the modulation scheme. It is hard to determine which modulations scheme is better than the other, it would depend on the network working area conditions. Also, is important to remark that the complexity of a 64 QAM or 16 QAM system is considerably higher than a BPSK or QPSK system, this impacts directly in the cost of the network implementation.

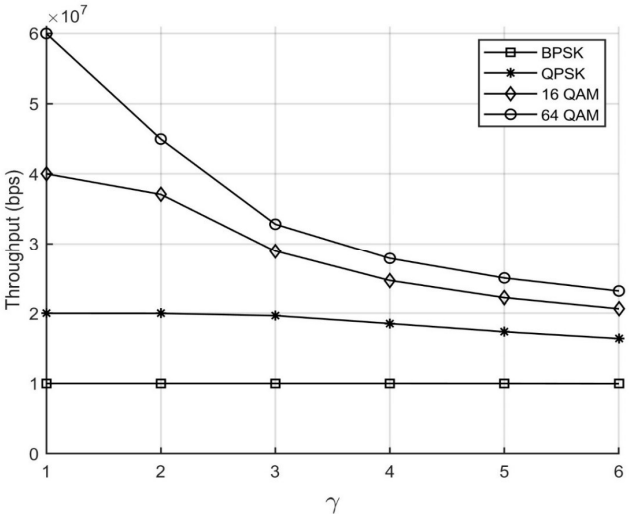


Fig. 4. Throughput Behavior for Node 4.

We can say that in general terms 64 QAM offers the best behavior when talking about channel usage or throughput compared to 16 QAM, QPSK and BPSK, but this would not be enough to determine which scheme is better to implement in a real-world scenario, it would be necessary to analyze other network and environmental metrics, also, a system supporting different modulation schemes can be possible.

Comparing the results obtained and the data report in other works for example in [7] a vanet is evaluated and throughput results are reported, the main difference is the coverage area simulated, in [7] the network cell has a 1km radios, obtaining significantly less bandwidth usage, for 64 QAM the maximum throughput obtained is 20.6 Mbps while in this work que maximum throughput for the same modulation is 60 Mbps. This confirms what we said before, the mobility and positioning affect considerably the network performance.

References

1. Brito-Gómez, J.: Evolución de las redes móviles hasta hoy en día y el impacto de la red móvil de quinta generación. Revista ReDTis, vol. 3, no. 3 (2019) http://www.redtis.org/index.php/Redtis/article/view/36
2. VIU Internacional: Evolución de la red de comunicación móvil, del 1G al 5G. Universidad Internacional de Valencia (2018) https://www.universidadviu.com/int/actualidad/nuestros-expertos/evolucion-de-la-red-de-comunicacion-movil-del-1g-al-5g
3. García-Santiago, A.: Diseño e implementación de una red de comunicaciones para el control de formación de vehículos aéreos no tripulados. Benemérita Universidad Autónoma de Puebla (2018) https://hdl.handle.net/20.500.12371/8205
4. Paul, A., Chilamkurti, N., Daniel, A., Rho, S.: Evaluation of vehicular network models. Intelligent Vehicular Networks and Communications, pp. 77–112 (2017) doi: 10.1016/B978-0-12-809266-8.00004-1

5. ITU: Comitted to connecting the world: Network dimensioning and configuration. Unión Internacional de Telecomunicaciones (2001) https://www.itu.int/itudoc/itud/dept/psp/ssb/mpg/ch07.pdf
6. Pióro, M., Medhi, D.: Routing, flow, and capacity design in communication and computer networks. The Morgan Kaufmann Series in Networking, pp. 679–711(2004) doi: 10.1016/B978-012557189-0/50024-X
7. Villegas-Hernández J. C.: Diseño de una plataforma de simulación para la evaluación del desempeño de un sistema celular OFDMA con estaciones base móviles. Benemérita Universidad Autónoma de Puebla (2020)
8. Castañeda-Camacho, J., Rios, C., Lara-Rodriguez, D.: Reverse link erlang capacity of multiclass CDMA cellular system considering nonideal antenna sectorization. IEEE Transactions on Vehicular Technology, vol. 52, no. 6, pp. 1476–1488 (2003) doi: 10.1109/TVT.2003.816629
9. Castañeda-Camacho, J., Sánchez-Mendoza A., Hernández-Rodríguez, D., Saviñón-López, R., Rodríguez-Domínguez, A. M., Maya-Rueda, S. E., Mino-Aguilar, G.: 5G downlink power quality evaluation for 3D environments. In: Artículos del Congreso Internacional de Investigación Academia Journals, vol. 13, no. 10, pp. 455-460 (2021)

ISSN 1870-4069

Proceso de calibración de sonda utilizada en la detección del nivel de potencial de hidrógeno en un sistema recolector de datos IoT para cultivos hidropónicos

Nicolás Quiroz-Hernández[1], Luis Efraín López-García[1], Antonio Martínez-Ruiz[2], Rodrigo Lucio Maya-Ramírez[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Electrónica, México

[2] Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, México

luis-efrain@hotmail.com

Resumen. Las técnicas de cultivo hidropónico tienen gran popularidad, debido a un mejor aprovechamiento del espacio y del agua. Este tipo de cultivo hace uso de soluciones nutritivas para realizar la nutrición de las plantas, esta debe ser controlada para obtener los mejores rendimientos. El pH al ser una característica importante se debe controlar su nivel en la solución nutritiva ya que este afecta la solubilidad de los nutrientes. Para monitorizar el pH se suelen utilizar sensores dedicados a través de sistemas recolectores de datos que usan sondas para cada característica de la solución nutritiva, estas sondas deben mantenerse calibradas para funcionar de forma adecuada. En este trabajo se muestra el proceso de calibración de una sonda de pH implementado para un sistema recolector de datos IoT. La calibración de la sonda se realiza utilizando dos puntos de calibración con soluciones que tengan pH conocido, de esta forma se obtiene una recta de calibración la cual es utilizada para mediciones que se muestran como resultado de este trabajo.

Palabras clave: Hidroponía, pH, calibración.

Probe Calibration Process Used in the Detection of the Hydrogen Potential Level in an IoT Data Collection System for Hydroponic Crops

Abstract. Hydroponic cultivation techniques are very popular, due to a better use of space and water. This type of cultivation makes use of nutritive solutions to carry out the nutrition of the plants, this must be controlled to obtain the best yields. The pH, being an important characteristic, its level in the nutrient solution must be controlled since it affects the solubility of the nutrients. To monitor pH, dedicated sensors are often used through data collection systems that use probes for each characteristic of the nutrient solution; these probes must be kept

calibrated to function properly. This paper shows the calibration process of a pH probe implemented for an IoT data collection system. The calibration of the probe is carried out using two calibration points with solutions that have a known pH, in this way a calibration line is obtained which is used for measurements that are shown as a result of this work.

Keywords: Hydroponics, pH, calibration.

1. Introducción

El cultivo hidropónico es una de las formas más eficientes de cultivar. La tasa de crecimiento de las plantas en un sistema hidropónico es mucho mayor que en un sistema basado en el suelo [1]. La hidroponía es un tipo de cultivo que no utiliza suelo para el cultivo de plantas. En su lugar, utiliza un medio inerte como grava, arena o lana de roca, que luego se riega con agua y nutrientes.

Como resultado, las raíces de la planta están expuestas al oxígeno y al agua, mientras que las hojas de la planta están expuestas a la luz. Esto permite una mejor absorción de nutrientes por parte de las raíces y tasas de crecimiento más rápidas en comparación con las plantas cultivadas en sistemas basados en el suelo [2]. Como el cultivo hidropónico es más eficiente que el suelo, se ha convertido en el método preferido por muchos agricultores.

El sistema de producción hidropónico se encuentra en constante crecimiento dentro del sector agroindustrial, así como en la investigación científica para la generación de nuevo conocimiento. Dentro de la agricultura de precisión se hace uso de sistemas tecnológicos enfocados a capturar datos dentro de los cultivos como mencionan en [3].

Con el crecimiento del sector hidropónico nace la necesidad de diseñar nuevos sistemas recolectores de datos enfocados a este tipo de cultivo, ya que según [4] la forma de trabajo en un cultivo hidropónico es diferente al cultivo tradicional, esto ocurre porque existen otras variables de importancia dentro de los mismos.

Existen muchos factores que afectan el crecimiento de un cultivo, entre ellos se tiene por ejemplo la genética de la planta en cuestión, la temperatura del ambiente a la que se encuentra sometida, la incidencia de luz sobre el cultivo, los nutrientes que se encuentran disponibles, etc [5].

En un cultivo hidropónico la solución nutritiva es muy importante, por ello se busca obtener la mayor cantidad de información referente a esta, se utilizan diferentes sensores y sondas para monitorizar las características que influyen en el crecimiento de la planta, como son, la temperatura, oxígeno disuelto, conductividad eléctrica y potencial de hidrógeno [6].

En el trabajo se habla sobre la importancia del pH para los cultivos hidropónicos, como funciona un sensor de pH y de qué forma se realiza la calibración de este. Se muestra el proceso seguido para realizar la calibración del sensor y sobre el sistema recolector de datos para observar la interconexión de los elementos. Por último, se muestran los resultados obtenidos, se da una comparativa entre esos resultados y los de otros trabajos, para terminar con unas conclusiones.

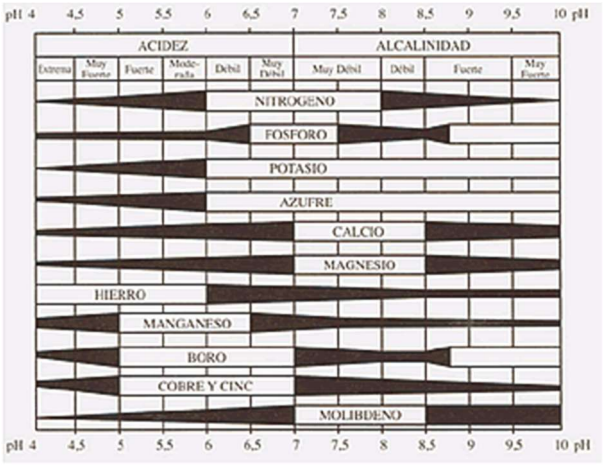


Fig. 1. Diagrama de Troug [9].

2. Importancia del pH en suelos y en soluciones nutritivas

La salud de una planta se ve afectada por diversos factores, siendo uno de estos las características con las que cuenta el suelo o la solución nutritiva utilizada, estas afectan la disponibilidad de macro y micronutrientes para el cultivo. El pH del suelo y las soluciones nutritivas es un factor determinante en la solubilidad de nutrientes, suelos que tienen un nivel de pH de 4.0 a 5.0 son considerados como ácidos [7]; los minerales como el aluminio y manganeso son más solubles en estos niveles lo cual puede ser toxico para el cultivo [8].

En suelos con un nivel de pH mayor a 8.0 o 9.0 la solubilidad de los nutrientes disminuye drásticamente por ende no se encuentran disponibles para que las plantas puedan asimilarlos. En la figura 1 se muestra un diagrama de Troug, una forma gráfica de observar la influencia del pH en la disponibilidad de los nutrientes. Se observan unas barras, dependiendo del grosor es la disponibilidad, a mayor grosor, mayor disponibilidad.

En el caso del fertirriego aplicado en cultivos de suelo o en hidroponía, el nivel de pH debe ser tal que permita que los nutrientes se disuelvan en su totalidad sin dañar el sistema radicular del cultivo, de esta forma se evita la formación de precipitados ya que estos pudieran causar obturaciones en los sistemas de riego. El rango ideal se encuentra entre 5.0 y 6.5. Por encima de 6.5 se da la formación de precipitados y por debajo de 5.0 se puede dar un daño en el sistema radicular del cultivo [10].

3. Sensor de potencial de hidrógeno

El nivel del pH se obtiene utilizando una sonda de medición, estas son conocidas por ser utilizadas en los pH-metros, a través de estos se realiza la medición con un método potenciométrico. Se establece que entre dos disoluciones con distinta H+ se genera una diferencia de potencial, esta diferencia de potencial se utiliza para determinar el nivel

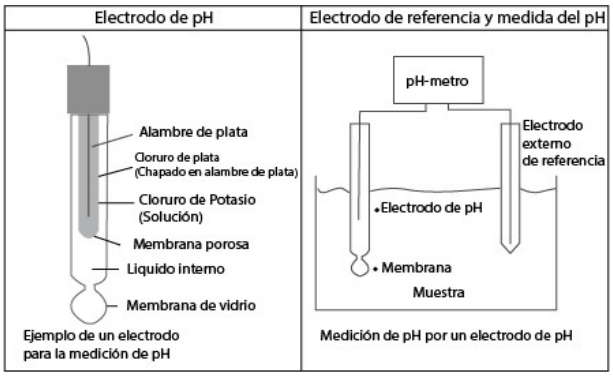


Fig. 2. Forma en la que se realiza la medición del pH [12].

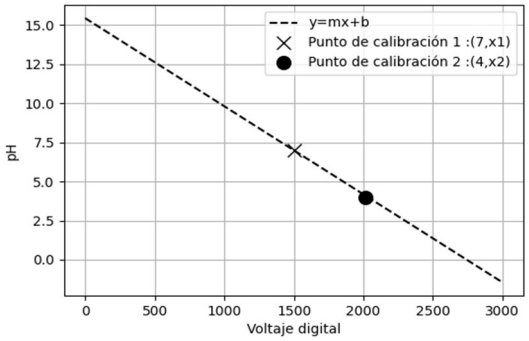


Fig. 3. Relación entre pH y voltaje digital.

de pH. La medida del pH es relativa, es decir se compara el pH de una muestra con el de una disolución con pH conocido como lo menciona [11]. Para realizar esto se utiliza un electrodo de pH como se puede observar en la figura 2, al entrar en contacto el electrodo con la disolución se establece un potencial a través de la membrana de vidrio que recubre el electrodo, el potencial generado varía según cambia el pH, para determinar el valor del pH se debe utilizar un electrodo de referencia cuyo potencial no varié.

La sonda utilizada en este trabajo lleva por nombre Analog pH Sensor, esta pertenece a la marca Gravity. La señal de salida que entrega esta sonda se encuentra filtrada utilizando un hardware dedicado, de esta forma se obtiene una señal analógica con baja fluctuación que se encuentra en el rango de 0 a 3.0 V. Esto permite un rango de detección de 0 a 14 en el nivel de pH, tiene una precisión de ±0.1 a 25° C y trabaja con una alimentación de 3.3 V a 5.5 V.

La señal que genera la sonda es leída utilizando un microcontrolador que tenga un convertidor analógico-digital, con esto se obtiene un valor digital que representa el voltaje medido por la sonda en ese instante de tiempo, con este valor es posible calcular de forma algebraica el nivel de pH, este cálculo se realiza utilizando la ecuación de la recta (1):

$$y = mx + b, \tag{1}$$

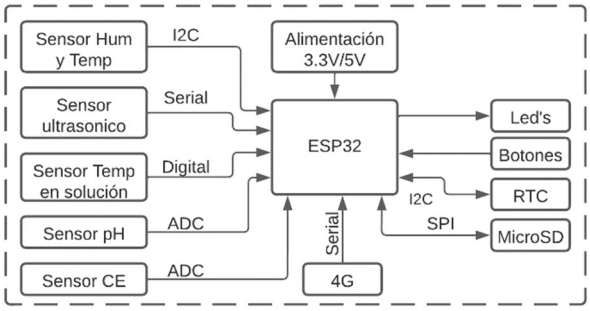


Fig. 4. Diagrama de bloques de sistema recolector de datos.

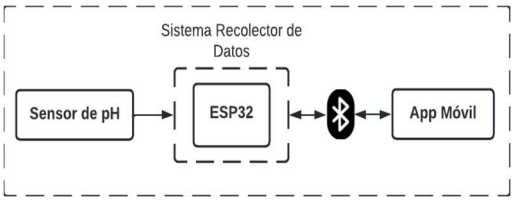


Fig. 5. Diagrama de bloques de sistema.

donde "**y**" representa el nivel de pH, "**x**" el valor digital de voltaje entregado por la sonda, "**m**" la pendiente y "**b**" la intercepción de la recta. Entonces para obtener el valor de la pendiente y de la intercepción es necesario contar con dos puntos que se encuentren dentro de la recta y así utilizar la ecuación (2) y (3) para calcular la pendiente y la intercepción respectivamente:

$$m = \frac{y_2 - y_1}{x_2 - x_1}, \tag{2}$$

$$b = y - mx. \tag{3}$$

En la figura 3 se muestra un ejemplo de recta que relaciona el pH con respecto a un voltaje, se tiene un primer punto de calibración donde se utiliza una solución con pH conocido de 7 y un segundo punto con pH conocido de 4. Al conocer estos dos puntos se pueden utilizar las ecuaciones (2) y (3) para obtener la pendiente y la intercepción, de esta forma se puede calcular el nivel del pH para cualquier solución que entregue un valor X.

4. Proceso de calibración

La sonda de pH se utiliza en un sistema recolector de datos IoT, en la figura 4 se muestra un diagrama de bloques de este sistema, en el cual se puede observar que cuenta un microcontrolador ESP32, este tiene integrado un modulo de comunicación Wi-Fi y Bluetooth, el sistema hace uso de cinco sensores para el cultivo hidropónico, cuenta con indicadores led, botones para reset y configuración, se puede utilizar una memoria SD para guardar los datos de forma local o a través del 4G y Wi-Fi enviar los datos al sistema de nube.

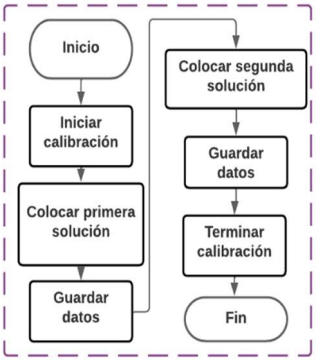


Fig. 6. Algoritmo de calibración.

El sistema cuenta con la capacidad de monitorear la humedad y temperatura relativa del ambiente, el pH, la conductividad eléctrica, el nivel y temperatura de la solución. Tres de estas sondas utilizan comunicación digital como I2C, UART y 1-WIRE, la otras dos entregan una señal analógica. Se miden estas variables ya que afectan procesos foto-respiratorios y enzimáticos de los cultivos.

Los datos son enviados a un sistema de nube desarrollado en AWS, se almacenan en una base de datos SQL y son desplegados en un panel a través de una aplicación WEB desarrollada utilizando Flask. La comunicación entre el sistema recolector de datos y la nube se realiza por Wi-Fi o por 4G haciendo uso del protocolo HTTP.

La calibración de la sonda de pH sigue el flujo que se muestra en la figura 5 donde con un diagrama de bloques se observa la conexión del sensor con el sistema recolector de datos y el uso de comunicación Bluetooth para dar enlace con una aplicación móvil, desde esta se configuran los parámetros de funcionamiento del sistema y se realiza la calibración de la sonda de pH.

La aplicación móvil a través de comunicación Bluetooth envía datos al microcontrolador ESP32 que se encuentra dentro del sistema recolector de datos, siendo así que cuando se inicia la calibración del sensor de pH, en la aplicación se le indica al usuario los pasos que debe seguir y al microcontrolador se le indica cuando debe realizar ciertas acciones.

Por dentro el microcontrolador al entrar en modo de calibración se encuentra tomando muestras con la sonda, con un comando se le indica que debe guardar el valor medido en ese instante, usando ciertos condicionales con valores precargados identifica la solución que se está utilizando, ya sea de 7.0 o de 4.0 y con un último comando se le indica el final del proceso de calibración.

La calibración se realiza siguiendo el algoritmo mostrado en la figura 6. Donde a través de la aplicación se le indica al sistema recolector de datos que se va a iniciar el proceso de calibración, el sistema espera a que se coloque una solución de calibración, en este caso se esperan soluciones con valor de 7.0 o de 4.0, al colocar la sonda dentro de la solución se detecta de forma automática cuál de las soluciones se utiliza y procede a guardar los datos que corresponderían a Y_1 y X_1 en la memoria del microcontrolador.

```
voltage:1359.16
temperature:22.7^C
pH:7.7936

>>>Enter PH Calibration Mode<<<
>>>Please put the probe into the 4.0 or 7.0 standard buffer solution<<<
```

Fig. 7. Inicio de Calibración.

```
voltage:1446.97
temperature:22.9^C
pH:7.2988

>>>Buffer Solution:7.0,Send EXITPH to Save and Exit<<<
```

Fig. 8. Detección de solución 7.0.

```
voltage:2005.30
temperature:23.0^C
pH:4.1391

>>>Buffer Solution:4.0,Send EXITPH to Save and Exit<<<
```

Fig. 9. Detección de solución 4.0.

A continuación, se coloca la segunda solución y de igual forma el sistema la detecta y almacena los valores de Y_2 y X_2 en memoria no volátil, con estos valores se hace uso de la ecuación (2) y (3) para obtener la pendiente y la intercepción, de esta forma el sistema se encuentra calibrado y utiliza la ecuación (1) para realizar futuras mediciones, por último, se indica al sistema que se ha terminado con el proceso de calibración.

5. Resultados

En la figura 7 se muestra la respuesta a través de una consola de comunicación serial con el microcontrolador, donde de inicio se observan las mediciones que está realizando el sistema de forma independiente, se muestra la temperatura ambiental y el pH detectado por la sonda, seguido de esto se le envía el comando al microcontrolador para indicar que se debe acceder al modo de configuración, donde el mismo microcontrolador responde que se accedió a este modo y solicita que se coloque la solución de 4.0 o 7.0.

La sonda previamente se encuentra inmersa en una solución de 3mol/L KCL que genera una capa de protección para su almacenamiento adecuado, al sacarla de esta solución se debe limpiar con agua destilada y posteriormente utilizar la sonda ya sin ningún problema.

Seguido de esto se coloca la sonda en una solución buffer para calibración con valor de 7.0, en la figura 8 se muestra el primer punto de calibración la sonda detecta la solución de 7.0 y guarda este punto de calibración en la memoria del microcontrolador.


```
voltage:1511.43
temperature:25.2^C
pH:7.0966
voltage:1510.62
temperature:25.2^C
pH:7.1012
voltage:1517.07
temperature:25.2^C
pH:7.0644
voltage:1510.62
temperature:25.2^C
pH:7.1012
voltage:1513.04
temperature:25.2^C
pH:7.0874
voltage:1512.23
temperature:25.2^C
pH:7.0920
```

(a) Medición en Solución de 7.0. (b) Lecturas de medición.

Fig. 2. Comprobación de calibración.

Tabla 1. Comparativa de mediciones.

Muestra	Valor real	Medición	Trabajo [13]	Trabajo [14]
1	7.0	7.0966	7.05	7.066
2	7.0	7.1012	7.07	7.021
3	7.0	7.0644	7.08	7.017
4	7.0	7.1012	7.1	6.995
5	7.0	7.0874	7.06	X
6	7.0	7.0920	7.08	X
	Error Promedio	0.09	0.07	0.02

Una vez realizado el primer punto de calibración es necesario limpiar nuevamente la sonda de pH utilizando agua destilada para eliminar los restos de la solución de 7.0 y de esta forma no se afecte el segundo punto de calibración. En la figura 9 se muestra la detección de la solución de 4.0 lo que da paso a que el sistema almacene en memoria no volátil este segundo punto de calibración.

Con esto se da por concluida la calibración y la sonda ya puede ser utilizada para realizar la medición del pH en distintas soluciones, ya que con estos puntos de calibración se da el ajuste de la recta y=mx+b que modela el comportamiento del nivel de pH con respecto a un voltaje.

Para comprobar que la calibración se dio de forma adecuada, se realiza una medición utilizando una solución con pH conocido, en este caso de 7.0. En las lecturas de la sonda se espera obtener mediciones de +- 0.1 a 25º.

En la figura 10 (a) se muestra la sonda colocada en la solución de 7.0 y en la figura 10 (b) las lecturas obtenidas. Se puede observar que se está teniendo una medición correcta dentro del margen de error establecido por el fabricante.

En [13] se hace uso de un sensor de pH en cultivos hidropónicos y en [14] para determinar la calidad del agua destilada con respecto a la norma ISO 17025/2005. Los autores exponen mediciones con respecto a una solución de 7.0 de pH, estas mediciones se exponen en la tabla 1, donde se comparan con las obtenidas en este trabajo. Las mediciones de este trabajo reportan mayor número de decimales de resolución en comparación a los otros dos trabajos.

El error promedio absoluto para las mediciones de este trabajo se encuentra dentro del rango que maneja el fabricante, de esta forma se observa que la calibración funciona de forma efectiva, en comparativa el trabajo [14] tiene un error promedio absoluto menor, esto se debe equipo que se está utilizando y la calidad de sonda utilizada, ya que el trabajo [14] busca cumplir los requisitos de una norma ISO.

6. Conclusiones

El éxito de los cultivos hidropónicos en gran parte se debe al manejo de una buena solución nutritiva, para ello es importante mantener en un nivel adecuado el pH, esto con el objetivo de tener una buena solubilidad de los nutrientes para que las plantas a través del sistema radicular puedan asimilarlos sin ningún problema, así mismo al tener una buena solubilidad de los nutrientes se evitan los precipitados y con ello las obturaciones de los sistemas de recirculación o de riego.

El proceso de calibración abordado en este trabajo permite que los elementos de medición utilizados en un sistema recolector de datos puedan entregar una medición adecuada. Por lo tanto, la sonda de medición de pH realiza mediciones dentro de un rango óptimo para una solución nutritiva de cultivos hidropónicos.

Con este proceso de calibración se le da al usuario la capacidad de hacer una calibración óptima utilizando soluciones buffer para tener dos puntos de calibración que permiten que las mediciones de la sonda de pH puedan mantener la fiabilidad y den mediciones acertadas. La calibración y mediciones obtenidas en este trabajo se han comparado con otros trabajos en la literatura y se muestra que lo propuesto cuenta con una mayor resolución y una precisión dentro del margen que indica el fabricante.

Referencias

1. Beltrano, J., Giménez, D. O.: Cultivo en hidroponía. Editorial de la Universidad de La Plata (EDULP) (2015) doi: 10.35537/10915/46752
2. Hernández, C. J., Hernández, J. L: Valoración productiva de lechuga hidropónica con la técnica de película de nutrientes (nft). Naturaleza y Desarrollo, vol. 3, no. 1, pp.11–16 (2005)
3. García, E., Flego, F.: Agricultura de precisión. Revista Ciencia y Tecnología, Tecnología Agropecuaria, pp. 89–116 (2008) https://www.maquinac.com/wp-content/uploads/2015/07/Agricultura-de-Precision-Universidad-de-Palermo.pdf
4. Alveal, M. A., Campos, K. C.: Estudio comparativo de sistemas de riego hidropónico y por goteo. Universidad del Bío-Bío (2014)
5. Brenes-Peralta, L. P., Jiménez-Morales, M. F.: Manual de producción hidropónica para hortalizas de hoja en sistemas nft (nutrient film technique). Tecnológico de Costa Rica (2014) https://hdl.handle.net/2238/6581

6. Carrasco, G., Ramírez, P., Vogel, H.: Efecto de la conductividad eléctrica de la solución nutritiva sobre el rendimiento y contenido de aceite esencial en albahaca cultivada en NFT. Universidad de Talca, Facultad de Ciencias Agrarias, Idesia (Arica), vol. 25, pp. 59–62 (2007)
7. Rivera, E., Sánchez, M., Domínguez, H.: pH como factor de crecimiento en plantas. Revista de iniciación científica, vol. 4, no. 2 (2018) doi: 10.33412/rev-ric.v4.0.1829
8. Leal-Ayala, O. G.: Rango de ph óptimo para el desarrollo de tomate (solanum lycopersicum l) y tilapia (oreochromis niloticus) en acuaponía. Master's thesis (2017)
9. Durán, J. M., Retamal, N., Moratiel, R.: pH: Concepto, medida y aplicaciones en agricultura y medioambiente. Escuela Técnica Superior de Ingenieros Agrónomos, Universidad Politécnica de Madrid (2023) https://www.infoagro.com/abonos/pH_informacion.htm
10. Virgen-Carvajal, J. M.: Procesamiento de señal ph para control por MC de un sistema de fertirriego en cultivos hidropónicos. Universidad de los Andes (2019) http://hdl.handle.net/1992/44452
11. Martínez, A. C., Ventura, I. D.: Pelargonidina extraída del rábano como sustituto de indicadores de ph ácido-base de origen sintético. Portal de la Ciencia, no. 10, pp. 93–104 (2016) doi: 10.5377/pc.v10i0.3012
12. Ayrton, N. P.: Sistema automatizado de corrección de pH para piscina. Universidad Tecnológica Nacional, Facultad Regional Villa María (2020)
13. Pacuar, E. S.: Medición de pH y conductividad eléctrica para el control de un sistema hidropónico NFT. Universidad Técnica del norte (2020) http://repositorio.utn.edu.ec/handle/123456789/10622
14. Aruquipa, C. J.: Calificación del conductímetro y pH-metro utilizado en la determinación de la calidad del agua destilada, para cumplir los requisitos de la norma ISO17025/2005. Universidad Mayor de San Andrés (2018) https://repositorio.umsa.bo/xmlui/handle/123456789/18784

ISSN 1870-4069

LoRaWAN Downlink Power Quality Evaluation for 3D Environments

Daniel Hernández Rodríguez, Josefina Castaneda Camacho,
German Ardul Muñoz Hernández, Gerardo Mino Aguilar

Benemérita Universidad Autónoma de Puebla,
Mexico

daniel.hernandezrodriguez@viep.com.mx,
josefinacastaneda@yahoo.com.mx,
{germanardul.munoz, gerardo.mino}@correo.buap.mx

Abstract. The exponential growth of the number of mobile devices has improved the development of technologies adapted to the new environmental requirements, Low Power Wide Area Networks have become an essential area of study and investigation due to its diverse types of applications, including the IoT (Internet-of-Things) environments, sensors monitoring, medical applications, etc. There are some tools for the analysis of the implementation of this technology that helps to identify its advantages and disadvantages within its implementation parameters. This paper presents a downlink power quality evaluation of the received signal in a 3D environment using the LoRaWAN parameters. Two propagation models log-distance and Okumura-Hata are included to measure the outage probability.

Keywords: LPWAN, LoRaWAN, 3D environment, network evaluation.

1 Introduction

Low Power Wide Area Networks (LPWAN) are an important type of wireless communications that has been growing up in the last few years due to its applications and advantages. This network technology offers an improvement on the covered area (10 to 40 km in rural areas and 1 to 5 km in urban areas), also its easy scalability at a low cost [1, 2].

Several companies are working with LPWAN networks such as the Institute of Electrical and Electronics Engineers (IEEE), the European Telecommunication Standards Institute (ETSI), 3rd Generation Partnership Project (3GPP), the Internet Engineering Task Force (IETF), and LoRa Alliance. These companies developed their standards. SigFox, LoRaWAN, and NB-IoT, offering efficient solutions to connect multiple smart devices [3, 4]. In a previous paper was presented an outage probability analysis for a 3D environment for some technologies.

This paper improves that analysis for LPWAN networks, using the requirements and parameters of the LoRaWAN technology created by LoRa Alliance [5].

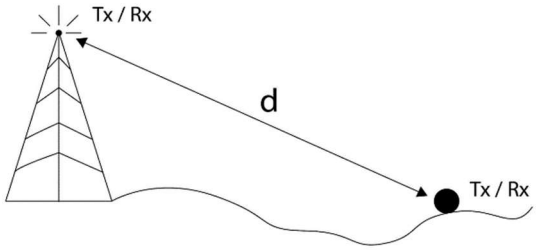


Fig.1. Scheme of the distance between the antenna and the end device.

When a network is being analyzed there are some affectations that must be included, one of those is the propagation losses that are generated in the wireless environment and it is caused by the distance of the devices, the antenna height, the devices height, and some other environment interferences.

This could be done by using some mathematical models that represent these affectations. Most of the LoRaWAN networks simulation tools, such as LoRaSIM uses the log-distance propagation model, but this is a bad approximation for a wide area analysis because it becomes inaccurate at long distances, a more realistic model for these kinds of networks is the Okumura-Hata model which is analyzed further in this paper [6].

This work is structured as follows. Second section presents the mathematical model of our system. Third section includes the simulation set-up and results, and finally in the last section is presented the conclusions.

2 Mathematical Model

In our simulation the coexistence of mobile and fixed end devices with a defined and undefined trajectory is analyzed. Inside the simulation the coexistence of mobile and fixed end devices is analyzed. Mobile devices movement is set on randomly in all three axes X, Y and Z. The analysis implies the received power evaluation considering the shadowing effects, distance losses using the log-distance and Okumura-Hata models, antenna gain (as a function of the user position) and 3D end devices distribution. The received power is expressed in the equation (1):

$$P_{RxM} = \frac{P_{Tx} \cdot G_{Tx} \cdot G_{Rx}}{d^{\mu}} \cdot 10^{\frac{\zeta}{10}}, \tag{1}$$

where P_{tx} is the transmission power, G_{Tx} and G_{Rx} are the antenna gains for the transmitter and the receiver respectively, d is the distance between the antenna and the end device, μ is the propagation losses exponent and ζ is the characteristic Gaussian random variable of the log-normal distribution which models the shadowing effects as shown in Fig. 1.

The transmitted antenna gain is considered as a parabolic function which depends on the angle $\phi_{l,n}$ between the end device and the antenna, this function is expressed in equation (2):

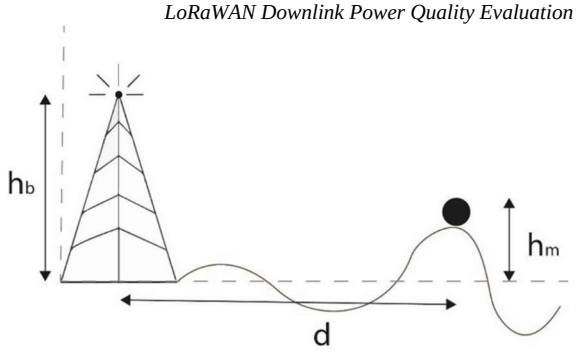


Fig.2. Illustration of the parameters used in the Okumura-Hata model.

Table 1. Maximum and minimum values for Okumura-Hata parameters.

Parameter	Symbol	Minimum value	Maximum value
Frequency	f	150 MHz	1500 MHz
Antenna height	h_b	30 m	200 m
Device height	h_m	1 m	10 m

$$G_{tx}(\phi_{l,n}) = \begin{cases} 1 - \frac{(1-q}{(\frac{\pi}{3})^2} \cdot \phi_{l,n}^2 \; si \; |\phi_{l,n} \leq \sqrt{\frac{1-p}{1-q}\frac{\pi}{3}} \\ \\ p \qquad\qquad\qquad si \;\; \phi_{l,n} > \sqrt{\frac{1-p}{1-q}\frac{\pi}{3}} \end{cases}, \tag{2}$$

where q represents the gain level of the antenna at a 60^o sector and p is the mean normalized gain of the side lobes [7].

In this simulation it has been evaluated the probability of the received power to be higher than the threshold given by the sensibility to guarantee the recovery of the signal and the link quality. This probability is given by the equation (3):

$$P(P_{rxM} > P_{min}) = \int_{P_{min}}^{\infty} f_{P-rxM}(x)dx \; . \tag{3}$$

Solving equation (3), we have the equation (4):

$$P(P_{rxM} > P_{min}) = \frac{1}{\sqrt{2\pi}} \int_{\frac{P_{min}-m_p}{\sigma_P}}^{\infty} e^{-\frac{u^2}{2}} du = Q(\frac{P_{min} - m_P}{\sigma_P}) \; , \tag{4}$$

where, $m_P = 10log(\frac{P_{tx}G_{tx}G_{rx}}{d^\mu}) + m_\zeta$
$\sigma_P^2 = E\{P_{rxM}dB^2\} - E^2\{P_{rxM}dB\} = \sigma_\zeta$.

The Okumura-Hata is based on empirical data that could be used to model the mobile propagation signals on rural and urban areas as a function of correction factor, the antenna height, end devices height, and the frequency [8]. The equation (7) describes the behavior of the Maximum Path Loss (MPL):

$$MPL_{dB} = A + B \cdot log(d) + C, \tag{5}$$

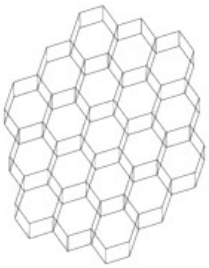


Fig.3. Hexagonal cells region.

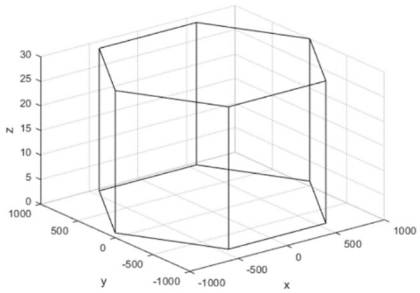


Fig. 4. Simulation 3D environment.

where, A, B, and C is described as follows:

$$A = 69.55 + 26.16log(f_c) - 13.82log(h_b) - a(h_m)$$

$$B = 44.9 - 6.55log(h_b)$$

$$C = 0.$$

C depends on the chosen environment, in this case, for metropolitan areas, f_c represents the central frequency given in MHz, d represents the distance in kilometers, h_b represents the antenna height or gateway in meters, h_m is the mobile height in meters and $a(h_m)$ is a function that represents the correction factor due to the antenna height and depends on the frequency of the communication, this function is given by the equation (6) [9]:

$$a(h_m) = \begin{cases} 8.29(log(1.54h_m)^2) - 1.1 \quad para\ f_c \leq 200MHz \\ 3.2(log(11.75h_m)^2) - 4.97\ para\ f_c \geq 400MHz \end{cases}. \tag{6}$$

The Fig. 2 shows graphically the parameters used to calculate the values of the Okumura-Hata model.

In the Table 1 are shown the maximum and minimum parameters that could be analyzed using Okumura - Hata model. Due it values, LoRaWAN Network could be analyzed with this model.

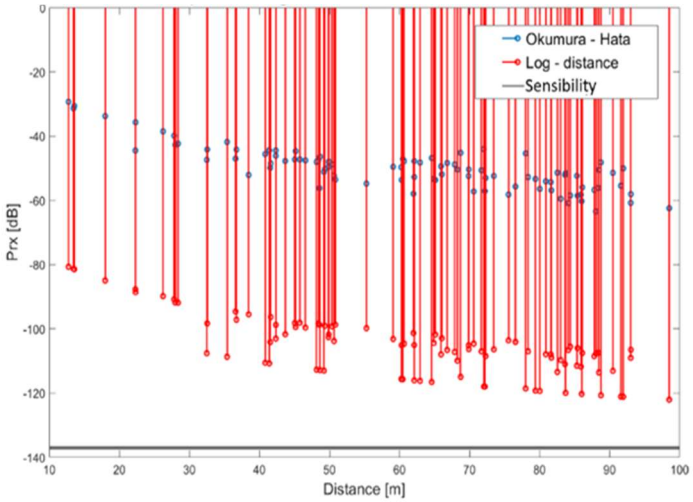


Fig.6. Received power for Log-Distance and Okumura-Hata for mobile users in a 100 meters radius.

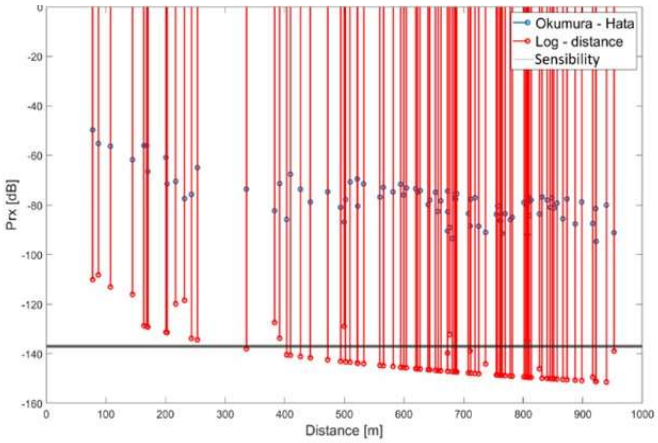


Fig.7. Received power for Log-Distance and Okumura-Hata for mobile users in a 1000 meters radius.

2.1 LoRaWAN Network

For a LoRaWAN simulation some parameters must be considered such as the frequency and bandwidth which depends on the unlicensed Industrial, Scientific and Medical (ISM) bands regulations for different areas, in US the accepted frequency is from 902 MHz to 928 MHz and the bandwidth is 125 kHz or 250 kHz, the spreading factor that defines the number of bits that could be transmitted on each symbol that could take values from 7 to 12, number of preambles that are used for the synchronization of the signal, the payload, the power of the transmitted signal and the sensibility that defines the minimum power level that could be received successfully.

Table 2. Maximum and minimum values for Okumura-Hata parameters.

Parameter	Value
Number of users	10-1000
Radius	100 - 1000 m
Frequency	902 MHz
Spreading Factor	12
Bandwidth	125 kHz
Transmission power	25.1mW / 14dB
Sensibility	1.98e-14W

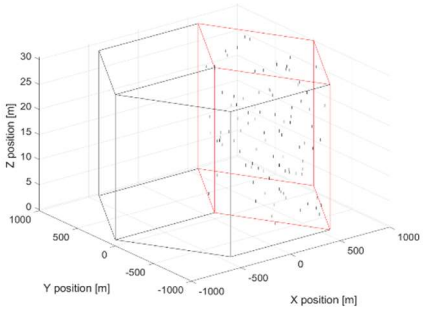


Fig.5. 3D environment for 100 users inside a radius of 1 km.

3 Simulation Set-Up

3.1 Simulation Scenario

For the simulation scenario we used the model developed in a previous paper which consider an environment divided in microcellular cells modeled as hexagonal prisms as shown in Fig. 3 [5].

The simulation occurs in a single cell with an adjustable radius, frequency, antenna height and device height. Considering a 120 degrees sectorization remarked in red color. This sector matches with the antenna propagation pattern placed in the center of the cell (black dot); this scenario is shown in Fig. 4 [5].

Once the simulation environment was set, several experiments were run, all of them using the mathematical model presented before. The value of the different parameters used are shown in the Table 2.

4 Results

The Fig. 5 represent the 3D environment with 100 users, the relation between mobile and fix users was randomly chosen within a radius of 1 km.

A comparison between both path loss models (Okumura-Hata in blue and log-distance in red) was done using different ratios, in the Fig. 6 is shown the behavior of

Table 3. Results for LoRaWAN using log-distance model.

Users	Events	Ratio [m]	Mean Prx [W]	Outage probability
10	1000	100	1.16e-9	1.9%
50	1000	100	4.93e-9	1.1%
100	1000	100	3.10e-8	2.5%
1000	1000	100	3.26e-7	1.9%
100	1000	200	8.20e-10	41.1%
100	1000	400	4.48e-12	82.3%
100	1000	500	3.16e-10	89.7%
100	1000	1000	1.54e-13	97.7%

Table 4. Results for LoRaWAN using Okumura-Hata model.

Users	Events	Ratio [m]	Mean Prx [W]	Outage probability
10	1000	100	4.70e-21	0%
50	1000	100	1.79e-19	0%
1000	1000	100	1.92e-20	0.01%
100	1000	200	6.24e-22	1%
100	1000	400	1.43e-23	9%
100	1000	500	2.66e-24	22%
100	1000	1000	2.31e-24	46%

both models at a 100 meters ratio, the black line represents the sensitivity of the system, in this case the result of both models is above the minimum power, therefore the signal will be correctly received.

If the ratio grown the outage probability for both models grown as well, for the log-distance model if the ratio is bigger than 300 meters the losses will be greater than the sensibility, in the Okumura-Hata model the losses at 300 meters are approximately 80 dB which is further from the sensibility (137 dB) as shown in Fig. 7.

The analysis which includes the log-distance and the Okumura-Hata models is presented in Table 3 for the log-distance and in the Table 4 for the Okumura-Hata. These results shown that the Okumura-Hata model works better for long distances.

5 Conclusions

We have observed for the log-distance that the outage probability increases over the 40% when the distance is greater than 200 meters which means that is not a reliable technology for long distances, considering that LoRaWAN was created for LPWAN technologies which its name represent wide areas log-distance is not a model that represents its behavior, the opposite occurs with Okumura-Hata which outage probability at the same distance (200 meters) is 1%, therefore in the study of LPWAN networks should be used the Okumura-Hata model.

References

1. Mekki, K., Bajic, E., Chaxel, F., Meyer, F.: A comparative study of LPWAN technologies for large-scale IoT deployment. ICT Express, vol. 5, no. 1, pp. 1–7 (2019) doi: 10.1016/j.icte.2017.12.005
2. Kosari, A., Wentzloff, D. D.: Murs band for LPWAN applications. In: IEEE Topical Conference on Wireless Sensors and Sensor Networks, pp. 1–3 (2019) doi: 10.1109/wisnet.2019.8711814
3. Yasmin, R., Petajajarvi, J., Mikhaylov, K., Pouttu, A.: On the integration of LoRaWAN with the 5g test network. In: IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, pp. 1–6 (2017) doi: 10.1109/pimrc.2017.8292557
4. Mekki, K., Bajic, E., Chaxel, F., Meyer, F.: Overview of cellular LPWAN technologies for IoT deployment: SigFox, LoRaWAN, and NB-IoT. In: IEEE International Conference on Pervasive Computing and Communications Workshops (2018) doi: 10.1109/percomw.2018.8480255
5. Castañeda-Camacho, J., Sánchez-Mendoza, A., Hernández-Rodríguez, D, Saviñón-Lopez, R., Rodríguez-Domínguez, A. M., Maya-Rueda, S. E., Mino-Aguilar, G.: 5G downlink power quality evaluation for 3D environments. In: Congreso Internacional de Investigación Academia Journals Celaya 2021, vol. 13, no. 10, pp. 455–460 (2021)
6. Bor, M. C., Roedig, U., Voigt, T., Alonso, J. M.: Do LoRa low-power wide-area networks scale? In: Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 59–67 (2016) doi: 10.1145/2988287.2989163
7. Habbal, A., Goudar, S. I., Hassan, S.: A context-aware radio access technology selection mechanism in 5G mobile network for smart city applications. Journal of Network and Computer Applications, vol. 135, pp. 97–107 (2019) doi: 10.1016/j.jnca.2019.02.019
8. Lee, C.: Mobile communications engineering theory and applications. McGrawHill (1998)
9. Francisco, S., Pinho, P., Luis, M.: Improving LoRa network simulator for a more realistic approach on LoRaWAN. In: Telecoms Conference, pp. 1–6 (2021) doi: 10.1109/conftele50222.2021.9435570

ISSN 1870-4069

Sistema de expediente clínico electrónico basado en aprendizaje automático

Ricardo Arturo López Álvarez[1], María del Carmen Santiago Díaz[1], Gustavo Trinidad Rubín Linares[1], Yeiny Romero Hernández[1], Judith Pérez Marcial[1], Ana Claudia Zenteno Vázquez[1], Julio César Díaz Mendoza[2]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México

[2] Universidad Autónoma de Yucatán, México

richla.5@outlook.com, {maricarmen.santiago, gustavo.rubin, yeiny.romero, judith.perez, ana.zenteno}@correo.buap.mx

Resumen. En México como en muchos países debido a la alta densidad de población se requiere brindar atención médica y de servicios a la sociedad, sin embargo, no es posible contar con el personal humano suficiente, por ello se debe digitalizar muchos recursos físicos y generar soluciones informáticas que puedan brindar atención a la sociedad al menos en la primera instancia. En éste trabajo se propone la metodología para construir expedientes clínicos y su administración mediante recursos informáticos, así como su mejoramiento mediante técnicas de aprendizaje automático que nos brinden una respuesta más óptima y además que requiera menos recursos físicos.

Palabras Claves: Expediente clínico, historial médico, expediente electrónico.

Electronic Medical Record System based on Machine Learning

Abstract. In Mexico, as in many countries, due to the high population density, it is necessary to provide medical care and services to society, however, it is not possible to have sufficient human personnel, therefore many physical resources must be digitized and computer solutions generated. that can provide care to society at least in the first instance. This paper proposes the methodology to build clinical records and their administration through computer resources, as well as their improvement through machine learning techniques that provide us with a more optimal response and also require less physical resources.

Keywords: Clinical record, medical record, electronic record.

1. Introducción

En México existen 37.5 millones de personas en carencia de acceso a los servicios de salud. Una persona se encuentra en esta situación cuando no cuenta con ninguna afiliación a las instituciones públicas de seguridad social o privadas [1].

Además, la secretaria de salud del país informa que la tercera causa de muerte es la diabetes donde las personas mayores de 65 años ocupan el primer lugar en muertes por esta causa. También nos dice que el 75% de los mexicanos tienen sobrepeso, una tercera parte de la población sufre obesidad. Sobre la hipertensión estima que cerca de 30 millones de personas la padecen y aproximadamente el 46% no lo sabe [2, 3].

Para tratar de reducir esta problemática se han creado diferentes organizaciones no gubernamentales(ONG) con el propósito de brindar atención a la población más vulnerable o con mayor dificultad para acceder a este servicio.

Las ONG (entidades sin fines de lucro - ESFL) conformadas por asociaciones y fundaciones son entidades privadas organizadas con autonomía de decisión, de libre adhesión, con voluntarios, que producen bienes o servicios sin contraprestación o con una menor. Que tienen como objetivo principal el interés general y sus excedentes se reinvierten en cumplir la misión [4].

1.1. Fundación mujeres haciendo historia

"La fundación mujeres haciendo historia" busca ayudar a la población más vulnerable del estado de Puebla brindando diferentes tipos de ayuda a las comunidades. Tiene un enfoque principal en la salud, priorizando la atención y prevención de enfermedades del tipo hipertensión, obesidad y diabetes, así como una inspección general del estado médico actual de la persona. Pero no solo se limita a este tipo de servicio ya que también ejerce diferentes actividades y programas.

Dentro de este tipo de organizaciones tenemos varios restos internos de los cuales podemos resaltar los recursos limitados al realizar las acciones. Se trabajará conjuntamente en desarrollar un sistema de expediente médico electrónico que permita llevar el control de la información generada, con el propósito de ayudar a la toma de decisiones, seguimientos y programas posteriores que pueda realizar esta organización. Complementando con la herramienta de apoyo al diagnóstico para disminuir la carga de trabajo médico.

1.2. Expediente clínico electrónico (ECE)

El expediente clínico electrónico es un archivo digital de la historia medica que contiene antecedentes clínicos relevantes, notas médicas, estudios de laboratorio, información administrativa e información que ayude a tener un perfil completo del paciente. El cual tiene diferentes ventajas como: La velocidad de acceso a la información. El almacenaje: reduce espacio, personal y papelería. Seguridad, solo los usuarios autorizados tienen acceso. Accesibilidad, la consulta es más sencilla. Mejora la infraestructura. Mejora la eficacia. Flexibilidad, se adapta a las necesidades de la institución [6].

De la revisión de la normal mexicana y diferentes libros médicos [6, 7]. Se obtuvo que una institución médica del tamaño de un hospital debe contar como mínimo con 18

Tabla 1. Desglose de las 4 entidades principales.

Entidad principal	Entidades que la conforman	Entidades derivadas
Historia Clínica (Hist_clinic)	Interrogatorio, exploración física, estudios previos, diagnóstico y terapéutica	
Notas		Ingreso, evolución, referencia/traslado, interconsulta, preoperatorias, postoperatorias, preanestésicas, posanestésicas, urgencias y egreso.
Hojas		Enfermería, indicaciones, egreso, egreso voluntario, notificación.
Documentos		Consentimiento, servicios auxiliares

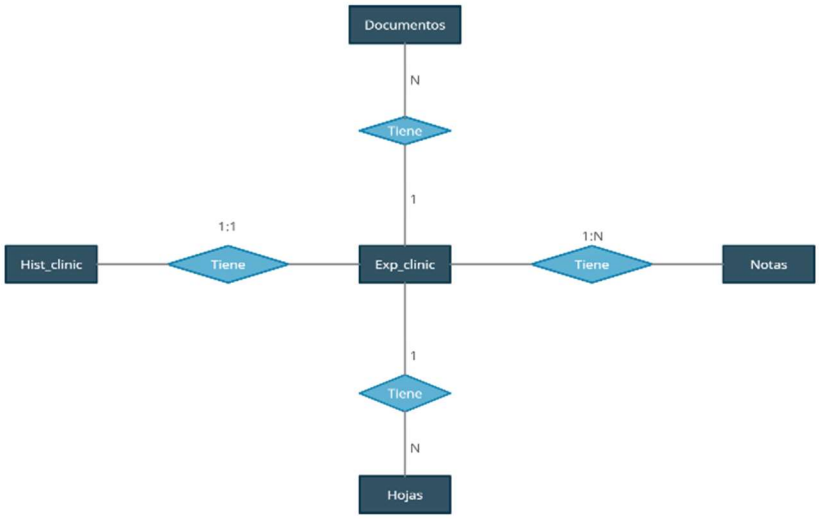


Fig. 1. Diagrama entidad relación del ECE principales entidades.

elementos para conformar un expediente los cuales se listan más adelante. Debido a esto nuestra base de datos constará de esos elementos.

1.3. Base de datos y modelo entidad relación

Una base de datos es un conjunto de información organizada o datos estructurados que se almacena de forma electrónica controlada por un sistema gestor de base de datos [8].

El modelo entidad relación es una técnica que tiene por objetivo la representación y definición de todos los datos que se introducen, almacenan, trasforman y producen dentro de un sistema de información [8].

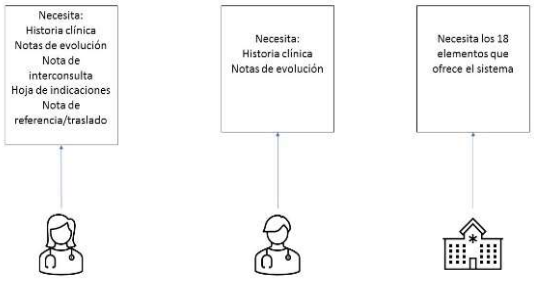


Fig. 2. Ejemplos de necesidades de los usuarios.

Fig. 3. Mapa funciones de sistemas.

Tener los datos estructurados en una base de datos nos permitirá usar técnicas de aprendizaje supervisado para desarrollar una herramienta.

1.4. Aprendizaje supervisado

El aprendizaje supervisado es una subrama de la inteligencia artificial y puede clasificarse en supervisado y no supervisado. En el aprendizaje supervisado los datos están etiquetados y se le pide al modelo hacer predicciones correctas y se le corrige cada que se equivoca, este proceso dura hasta que se alcance un buen nivel de precisión. En el aprendizaje no supervisado los datos no están etiquetados y el modelo se prepara deduciendo estructuras presentes en los datos.

El aprendizaje supervisado ha demostrado ser más efectivo para predicciones de ciertas enfermedades como diabetes, pero requiere de técnicas complementarias para equilibrar los datos y reducir la dimensionalidad seleccionando las entidades optimas, los modelos que han mostrado mejores resultados son los tipos árbol [9].

Tabla 2. Espacio de búsqueda para la creación del modelo.

Partes del espacio de búsqueda	Función	En el modelo
Espacio de datos	Contiene un conjunto finito de datos sobre el sistema y su entorno	Serán los síntomas y datos de entorno del paciente
Espacio de hipótesis	Contiene hipótesis (fallos) que puede tener el sistema	Esta parte serán las enfermedades
Espacio de reparaciones	Incluye acciones para recupera la funcionalidad	Ideas de terapéuticas a utilizar

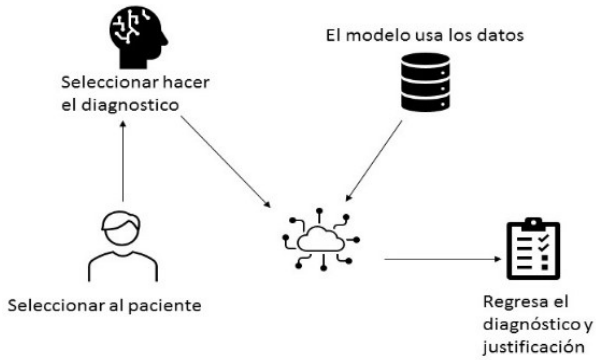


Fig.4. Diagrama de funcionamiento de la herramienta

2. Técnicas y métodos

2.1. Modelo de la base de datos

Buscamos que la fundación pueda guardar la información necesaria de los beneficiarios. Debido principalmente al enfoque en la salud, necesitamos una base de datos que funcioné como un ECE con el objetivo de mejorar el servicio, disminuir la carga de trabajo, la toma de decisiones y la gestión de la información.

Este apartado muestra las entidades de las cuales constará nuestro ECE y las relaciones que lo conforman. Una entidad es un objeto del mundo real del cual queremos guardar información, por cuestiones de claridad de imagen sólo se muestran las 4 entidades principales (véase la figura 1).

Estas cuatro entidades principales se conforman por más entidades que al final resultaran en los 18 elementos que se describen a continuación en las entidades derivadas (tabla 1).

Esta división en las entidades nos permite hacer flexible el sistema para adaptarlo a las necesidades y capacidades de diferentes tipos de establecimientos médicos,

permitiendo al usuario seleccionar las entidades que usará, obteniendo diferentes configuraciones de las entidades activas (fig.2). En el caso de la fundación haremos uso de la historia clínica, nota de evolución y documento de consentimiento.

2.2. Funciones del sistema

El sistema no solo será una herramienta de almacenamiento de información. Contará con diferentes funciones con las que pueda brindar mayor ayuda a los usuarios que de igual manera se pueden adecuar a sus necesidades (fig.2).

2.3. Herramienta de apoyo al diagnóstico

Se espera que el sistema pueda clasificar síntomas con diferentes enfermedades en base a la información del paciente radicando como una herramienta de apoyo al diagnóstico médico. Así como predecir y alertar tendencias a padecer enfermedades como diabetes, hipertensión obesidad y problemas debido al abuso de sustancias (alcohol, tabaco). No siempre se cuenta con gran personal de atención en organizaciones de índole no lucrativa.

Para esto haremos uso de técnicas de aprendizaje automático. Para explicar la función de los datos en la creación del modelo usaremos las bases de un sistema de diagnosis para hacer la explicación muy clara, tenemos un espacio de búsqueda conformado de un espacio de datos, un espacio de hipótesis y un espacio de reparaciones (véase tabla1).

El usuario seleccionará al paciente del cual desee un diagnóstico, el sistema le regresará la probabilidad de que sea una enfermedad haciendo uso de los datos o información contenidos en la historia clínica y en las notas de evolución si es que tiene. (véase figura 3).

3. Conclusiones

Se presentan las bases para la implementación de un expediente clínico que se implementará en poblaciones piloto y generará una base de datos que nos brinde información de las variables implícitas que requieren optimizarse o modificarse a fin de mejorar el desempeño de los algoritmos de aprendizaje. Aunque la base de datos inicial ya se generó con la población de prueba, se requiere mayor validación debido a la alta confidencialidad y riesgo de la información que se está procesando.

Referencias

1. CONEVAL: Nota técnica sobre la carencia por acceso a los servicios de salud 2018-2020 (2021) www.coneval.org.mx/Medicion/MP/Documents/MMP_2018_2020/Notas_pobreza_2020/Nota_tecnica_sobre_la_carencia_por_acceso_a_los_servicios_de_salud_2018_2020.pdf
2. Instituto de Salud para el Bienestar: Día mundial contra la obesidad (2022) www.gob.mx/insabi/articulos/dia-mundial-contra-la-obesidad-4-de-marzo?idiom=es

3. Secretaría de Salud: En México, más de 30 millones de personas padecen hipertensión arterial: Secretaría de Salud (2022) www.gob.mx/salud/prensa/238-en-mexico-mas-de-30-millones-de-personas-padecen-hipertension-arterial-secretaria-de-salud
4. Elechiguerra-Arrizabalaga, C., Corral-Lage, J., Maguregui Urionabarrenechea, M. L.: La gestión de asociaciones y fundaciones: Calidad y transparencia. Pirámide (2015)
5. Diario Oficial de la Federación: NORMA oficial mexicana NOM-004-SSA3-2012 (2012)
6. Ornelas-Aguirre, J. M.: El expediente clínico. Manual moderno (2013)
7. Keith-Stone, C., Humphries, R. L.: Diagnóstico y tratamiento en medicina de urgencias. LANGE, 7ma Edición (2013)
8. Mora-Rioja, Arturo: Bases de datos. Diseño y gestión (2014)
9. Fregoso-Aparicio, L., Noguez, J., Montesinos, L., García-García, J. A.: Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetology and Metabolic Syndrome, vol. 13, no. 1 (2021) doi: 10.1186/s13098-021-00767-9
10. Silaparasetty, N.: Machine learning concepts with python and the jupyter notebook environment (2020) doi: 10.1007/978-1-4842-5967-2

ISSN 1870-4069

Desarrollo de una metodología para control dinámico de motores con Machine Learning

Antonio Eduardo Álvarez Núñez[1], María del Carmen Santiago Díaz[1], Ana Claudia Zenteno Vázquez[1], María Catalina Rivera Morales[2], María Dolores Guevara Espinosa[2], Gustavo Trinidad Rubín Linares[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México

[2] Benemérita Universidad Autónoma de Puebla, Facultad de Ingeniería Química, México

{maricarmen.santiago, ana.zenteno, gustavo.rubin}@correo.buap.mx

Resumen. En robótica el problema de la precisión en el control de posición de los diferentes sistemas motrices requiere de la evaluación de múltiples variables en tiempo real a fin de no ocasionar una inestabilidad en el robot por el tiempo de procesamiento. En éste trabajo se propone una metodología basada en el aprendizaje automático de los sistemas motrices a fin de poder generar acciones de control con antelación a su medición y procesamiento, para lo cual se propone un sistema de segundo orden y su respuesta a perturbaciones como señales escalón y rampas.

Palabras clave: Aprendizaje automático, control dinámico, asistentes inteligentes.

Development of a Methodology for Dynamic Motor Control with Machine Learning

Abstract. In robotics, the problem of precision in the position control of the different motor systems requires the evaluation of multiple variables in real time in order not to cause instability in the robot due to processing time. In this paper, a methodology based on automatic learning of motor systems is proposed in order to be able to generate control actions in advance of their measurement and processing, for which a second-order system and its response to disturbances such as step signals are proposed. and ramps.

Keywords: Machine learning, dynamic control, intelligent assistants.

1. Introducción

Inteligencia artificial (IA) se refiere a sistemas o máquinas que imitan la inteligencia humana para realizar tareas y pueden mejorar iterativamente a partir de la información que recopilan. La IA se manifiesta de varias formas. Algunos ejemplos son:

- Los chatbots utilizan la IA para comprender más rápido los problemas de los clientes y proporcionar respuestas más eficientes.
- Los asistentes inteligentes utilizan la IA para analizar información crítica proveniente de grandes conjuntos de datos de texto libre para mejorar la programación.
- Los motores de recomendación pueden proporcionar recomendaciones automatizadas para programas de TV según los hábitos de visualización de los usuarios.

La IA se ha convertido en un término general para las aplicaciones que realizan tareas complejas que antes requerían aportes humanos, como la comunicación online con los clientes o jugar al ajedrez. El término a menudo se usa indistintamente con sus subcampos, que incluyen el aprendizaje automático y el aprendizaje profundo. Sin embargo, hay ciertas diferencias.

Por ejemplo, el machine learning se centra en la creación de sistemas que aprenden o mejoran su rendimiento en función de los datos que consumen. Es importante tener en cuenta que, aunque todo machine learning es IA, no toda la IA es machine learning [1].

1.1. Machine Learning

El Machine Learning, o aprendizaje automático, es una rama de la inteligencia artificial que permite que las máquinas aprendan sin estar programadas para este propósito específico. Una habilidad esencial para crear sistemas que no solo sean inteligentes, sino autónomos y capaces de identificar patrones en los datos para convertirlos en predicciones [2].

1.2. Tipos de Machine Learning

El aprendizaje automático clásico a menudo se clasifica según la forma en que un algoritmo aprende a ser más preciso en sus predicciones. Hay cuatro enfoques básicos: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi supervisado y aprendizaje por refuerzo. El tipo de algoritmo que los científicos de datos eligen usar depende del tipo de datos que quieran predecir [3].

- **Aprendizaje supervisado:** en este tipo de aprendizaje automático, los científicos de datos proporcionan algoritmos con datos de entrenamiento etiquetados y definen las variables que quieren que el algoritmo evalúe para las correlaciones. Se especifica tanto la entrada como la salida del algoritmo.

Tabla 1. Parámetros de la data base.

Tiempo de simulación	Muestreo	Cantidad de entradas escalón
10 s	100	50

Tabla 2. Planta con entradas seno.

	Con ruido	Sin ruido
Normalizado	Método regresión lineal MSE: 21426377140726341292575948 8.0000 k-Nearest Neighbors para regresión multi-output MSE: 27.1421 Decision Tree para regresión multi-output MSE: 89.9494 Red neuronal multi-output MSE: 2935.1144	Método regresión lineal MSE: 5.1602 k-Nearest Neighbors para regresión multi-output MSE: 27.1174 Decision Tree para regresión multi-output MSE: 141.6744 Red neuronal multi-output MSE: 2604.0644
No normalizado	Método regresión lineal MSE: 47303403797533668765859 84.0000 k-Nearest Neighbors para regresión multi-output MSE: 133.0015 Decision Tree para regresión multi-output MSE: 174.3104 Red neuronal multi-output MSE: 8.8519	Método regresión lineal MSE: 125.3017 k-Nearest Neighbors para regresión multi-output MSE: 132.9581 Decision Tree para regresión multi-output MSE: 174.3605 Red neuronal multi-output MSE: 5.4136

- **Aprendizaje no supervisado:** este tipo de aprendizaje automático implica algoritmos que se entrenan con datos no etiquetados. El algoritmo escanea a través de conjuntos de datos en busca de cualquier conexión significativa. Los datos con los que se entrenan los algoritmos, así como las predicciones o recomendaciones que generan, están predeterminados.
- **Aprendizaje semi supervisado:** este enfoque de aprendizaje automático implica una combinación de los dos tipos anteriores. Los científicos de datos pueden alimentar un algoritmo mayormente etiquetado como datos de entrenamiento, pero el modelo es libre de explorar los datos por sí mismo y desarrollar su propia comprensión del conjunto de datos.
- **Aprendizaje por refuerzo:** los científicos de datos suelen utilizar el aprendizaje por refuerzo para enseñar a una máquina a completar un proceso de varios pasos para el que existen reglas claramente definidas. Los científicos de datos programan un algoritmo para completar una tarea y le dan señales positivas o

negativas a medida que descubre cómo completar una tarea. Pero en su mayor parte, el algoritmo decide por sí mismo qué pasos tomar en el camino.

Los datos multivariados ocurren en una variedad de disciplinas, por ejemplo, en la investigación biomédica, las ciencias sociales, o la econometría. Se dice que los datos son multivariados si la respuesta no solo consta de una variable, sino de $d \geq 2$ variables de salida, digamos $Y \in R$. Entonces, a menudo estamos interesados en encontrar una relación funcional entre la salida Y y algunas características variables $X \in R$, es decir, queremos realizar un análisis de regresión multivariable.

A diferencia de la regresión múltiple uní variable ($d = 1$), que también incluye características múltiples de X, la regresión multivariable trata de especificar la relación de varias variables de resultado con X simultáneamente.

El objetivo de tales análisis multivariables es que la consideración de las posibles dependencias entre los resultados pueda conducir a procedimientos con mejor potencia (en caso de inferencia) o precisión (en caso de predicción) en comparación con análisis univariados separados [1].

2. Metodología

La regresión es una tarea de modelado predictivo que implica predecir una salida numérica dada alguna entrada.

Es diferente de las tareas de clasificación que implican predecir una etiqueta de clase.

Por lo general, una tarea de regresión implica predecir un solo valor numérico. Aunque, algunas tareas requieren predecir más de un valor numérico. Estas tareas se conocen como regresión de salida múltiple o regresión de salida múltiple para abreviar.

En la regresión de múltiples salidas, se requieren dos o más salidas para cada muestra de entrada y las salidas se requieren simultáneamente. La suposición es que las salidas son una función de las entradas [2].

Para que funcionen mejor muchos algoritmos de Machine Learning usados en Data Science, hay que normalizar las variables de entrada al algoritmo. Normalizar significa, en este caso, comprimir o extender los valores de la variable para que estén en un rango definido. Sin embargo, una mala aplicación de la normalización, o una elección descuidada del método de normalización puede arruinar tus datos, y con ello tu análisis [3].

Existen técnicas de normalización que comprime los datos de entrada entre unos límites empíricos (el máximo y el mínimo de la variable). Esto quiere decir que, si existe ruido, éste va a ser ampliado [3].

Para ver el efecto de normalizar los datos y el tener ruido en la señal que se medirá un futuro directamente desde los sensores del robot, se decidió hacer cuatro casos de estudio:

1. Implementar los algoritmos con el conjunto de datos sin normalizar y sin ruido.
2. Implementar los algoritmos con el conjunto de datos sin normalizar y con ruido.
3. Implementar los algoritmos con el conjunto de datos normalizados y sin ruido.
4. Implementar los algoritmos con el conjunto de datos sin normalizar y con ruido.

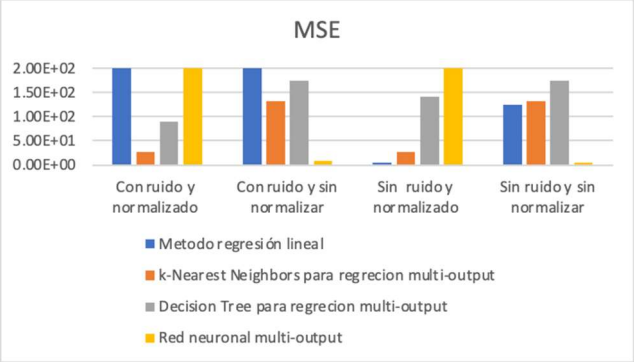


Fig. 1. Comparación de los errores cuadráticos medio.

Tabla 3. Comparación de la red neuronal con entrada escalón en las tres implementaciones.

	Con ruido y normalizado	**Con ruido y sin normalizar**	**Sin ruido y normalizado**	**Sin ruido y sin normalizar**
Red neuronal multi-output (1)	20215,8381	0,4571	18726,291	0,1747
Red neuronal multi-output (2)	2494159,312	0,5066	163.150	0,0981
Red neuronal multi-output (3)	284.105	0,0776	276104,4268	38912,2673

3. Resultados

A continuación, vemos los resultados, primero se dará un vistazo a los algoritmos con la planta recibiendo escalones unitarios y después recibiendo entradas seno.

3.1. Planta con entradas seno

Aquí la predicción es mejor cuando los datos no tienen ruido, y esta sin normalizar. También la red neuronal mejora su comportamiento cuando los datos no se normalizan.

En la tabla podemos ver que el error cuadrático medio (MSE) disminuye cuando los datos no contienen ruido. Por lo tanto, se recomienda que los datos obtenidos de los sensores pasen por un tratamiento de señal para reducir el ruido de estas. Por otro lado, aún no se puede concluir que normalizar los datos afecten la predicción de los modelos, ya que esto también se puede ver afectado por la cantidad de información con la que se entrenan los modelos. Por lo tanto, queda generar una base de datos más extensa para ver si los resultados mejoran.

3.2. Comparación de los resultados de las tres metodologías

En las Fig. 2 y 3 y en la tabla 4 se presentan los resultados de aplicación de las tres metodologías descritas anteriormente.

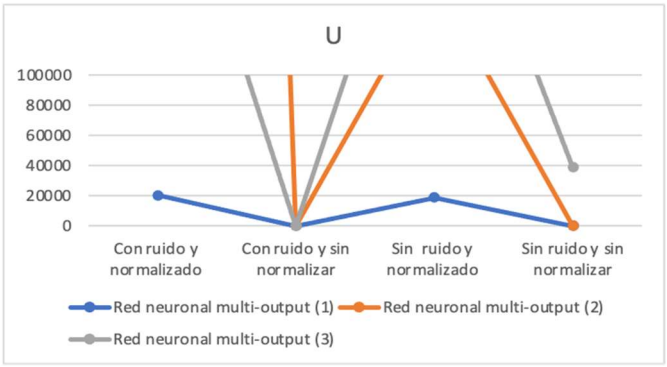


Fig. 2. Comportamiento de la red neuronal en las tres implementaciones.

Tabla 4. Comparación de la red neuronal con entrada escalón en las tres implementaciones.

	Con ruido y normalizado	**Con ruido y sin normalizar**	**Sin ruido y normalizado**	**Sin ruido y sin normalizar**
Red neuronal multi-output (1)	2935,1144	8,8519	2604,0644	5,4136
Red neuronal multi-output (2)	285180,3809	1478,1502	163.150	1105,6508
Red neuronal multi-output (3)	7.225	17,5880	13993,253	13,2334

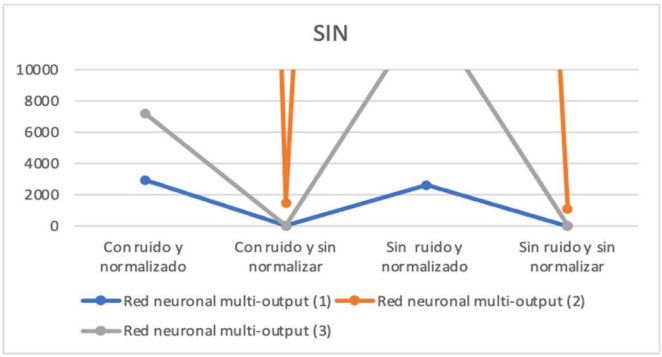


Fig. 3. Comportamiento de la red neuronal en las tres implementaciones

4. Conclusiones

Podemos observar que las predicciones mejoran en el caso donde los datos no son normalizados, también se observa que en la segunda metodología no hizo que las predicciones mejoraran.

Un punto importante por remarcar es que la red neuronal presenta mejor desempeño con la primera metodología, ya que se diseñó la red neuronal para ese caso, para las

siguientes dos metodologías se ocupó la misma red neuronal, los datos mejorarían si se hace una red neuronal para cada caso.

Para tener una red neuronal capaz de predecir cualquier entrada no solo la escalón y seno, es necesario entrenar la red con entradas aleatorias, este es la siguiente metodología que seguir.

References

1. Schmid, L., Gerharz, A., Groll, A., Pauly, M.: Machine learning for multi-output regression: when should a holistic multivariate approach be preferred over separate univariate ones? (2022) doi: 10.48550/ARXIV.2201.05340
2. Brownlee, J.: How to develop multi-output regression models with python. Machine Learning Mastery, Ensemble Learning (2020) machinelearningmastery.com/multi-output-regression-models-with-python/
3. Morante, S.: Precauciones a la hora de normalizar datos en data science. Telefónica Tech (2018) empresas.blogthinkbig.com/precauciones-la-hora-de-normalizar/

ISSN 1870-4069

Desarrollo de un Robot asistente para detección de Alzheimer

Juan Sebastián Orozco Van[1], María del Carmen Santiago Díaz[2], Judith Pérez Marcial[2], Ana Claudia Zenteno Vázquez[2], Yeiny Romero Hernández[2], Hermes Moreno Álvarez[3], María Catalina Rivera Morales[4], Gustavo Trinidad Rubín Linares[2]

[1] Universidad Autónoma de Occidente, Facultad de Ingeniería, Colombia

[2] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México

[3] Benemérita Universidad Autónoma de Puebla, Facultad de Ingeniería Química, México

[4] Universidad Autónoma de Chihuahua, Departamento de Ingeniería Aeroespacial, México

juan_s.orozco@uao.edu.co, {maricarmen.santiago, judith.perez, ana.zenteno, yeiny.romero, gustavo.rubin}@correo.buap.mx

Resumen. En la actualidad, se ha evaluado el potencial de la robótica como una herramienta valiosa para el apoyo del tratamiento de la demencia. El propósito de este estudio es realizar un primer acercamiento de la robótica hacia el monitoreo de anomalías del comportamiento del paciente. Para el desarrollo del proyecto, se ha realizado un modelo 3D de un robot usando el programa SolidWorks, se simulo el sistema de percepción inicial, y finalmente se desarrolló un algoritmo de navegación básico para iniciar la ejecución de las tareas de seguimiento del paciente con Alzheimer y medir de acuerdo a la evolución temporal y ejecución de sus actividades cotidianas.

Palabras clave: Robótica, monitoreo, demencia.

Development of an Assistant Robot for Alzheimer's Detection

Abstract. Currently, the potential of robotics as a valuable tool to support dementia treatment has been evaluated. The purpose of this study is to make a first approach of robotics towards the monitoring of patient behavior

abnormalities. For the development of the project, a 3D model of a robot has been made using the SolidWorks program, the initial perception system was simulated, and finally a basic navigation algorithm was developed to start the execution of the monitoring tasks of the Alzheimer's patient. and measure according to the temporal evolution and execution of their daily activities.

Keywords: Robotics, monitoring, dementia.

1. Introducción

El Alzheimer es un trastorno neurodegenerativo asociado con una deficiencia progresiva en memoria y habilidades cognitivas, así como la pérdida de habilidades para pensar. De hecho, es la forma más común de demencia, de la cual se le atribuye entre el 60% y 80% de casos [1, 2].

De acuerdo con las estadísticas actuales, alrededor de seis millones de americanos mayores de 65 años viven con Alzheimer, con una proyección de 13.8 millones para el 2060[1]. El panorama no es alentador, pues de acuerdo con la tasa de prevalencia global, cada tres segundos, una persona se enferma de Alzheimer [3].

Además, se estima que la población en la Tierra será de 11.2 billones para el 2100 [3]. Por lo tanto, la población de adultos veteranos será de dos billones aproximadamente antes de la mitad del siglo XXI, quienes son los que tienden a sufrir de este tipo de condiciones [3].

Por otro lado, el costo total estimado estadounidense del cuidado del Alzheimer era de aproximadamente 355 billones de dólares en 2021, y se espera que crezca a un trillón de dólares para 2050 [1]. Otra problemática es que no existe una medicación aprobada para un tratamiento la cura del Alzheimer ni para detener sus síntomas [2].

Ante esta situación, se propone el desarrollo de un robot de asistencia social como un apoyo personalizado de bajo costo para los médicos con respecto al monitoreo de la salud de pacientes mayores de 65 años. Hoy en día, los robots de asistencia social presentan una popularidad ascendente en el monitoreo de la salud de personas de la tercera edad, pues, en el futuro, serán una herramienta fundamental para el cuidado de ellos [4].

Con respecto al tratamiento de la demencia, los robots representan una manera efectiva de llevar a cabo el cuidado de los pacientes, porque pueden brindar atención, sin quejas ni fatiga, en tareas que requieran de alta repetibilidad [5]. Así mismo, representan servicios inteligentes y adaptables a las necesidades del paciente [6].

Además, se plantea que el robot ejecute un seguimiento autónomo del paciente en tiempo real. Este tipo de robot (robot seguidor de persona) busca obtener la posición de una persona líder, y seguirla continuamente mientras ejecuta una tarea cooperativa [7]. Cabe destacar que seguir a la persona líder no es la operación primaria, pero es necesario para tener éxito en la misma [8].

Para estos robots, los factores que determinan sus características son: el medio de operación, la selección de sensores, el modo de interacción (explicita, donde hay una interacción humano-robot directa, e implícito, donde es indirecta), la granularidad (la cantidad de robots y humanos involucrados en toda la operación), y el grado de autonomía [8].

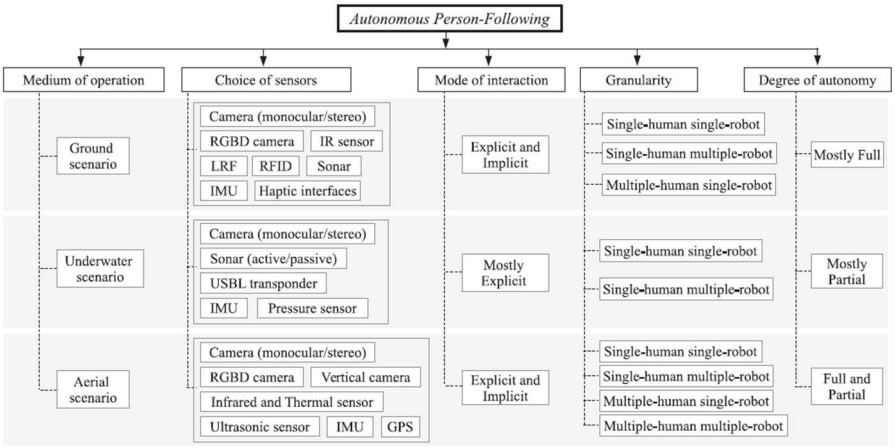


Fig. 1. Robots seguidores de personas y sus categorías.

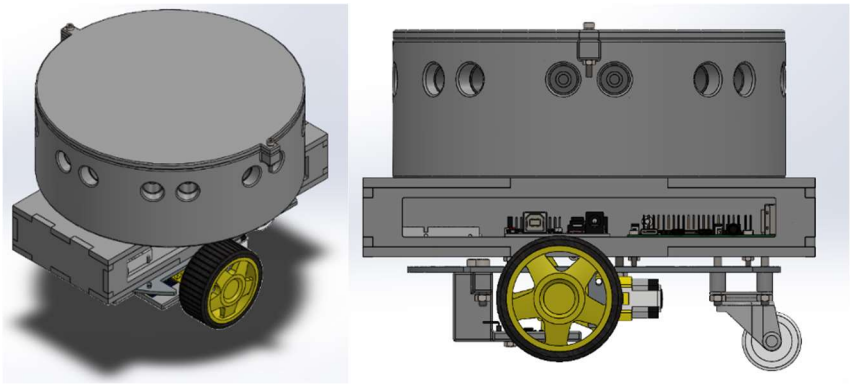


Fig. 2. Primer diseño de DETER.

A continuación, se presenta una figura con las categorías para robots seguidores de persona:

En éste trabajo se presenta el uso de la robótica y la inteligencia artificial como apoyo para el tratamiento de la demencia. Este estudio se realizó por el interés de explorar cómo la robótica y la inteligencia artificial puede respaldar a los médicos encargados sobre el estado de sus pacientes con respecto a las anomalías en su comportamiento.

Para el escenario propuesto, se plantea un robot terrestre (específicamente, diferencial) completamente autónomo, con un modo de interacción implícita, así como una granularidad de un humano y un robot.

La finalidad de esta investigación consiste en apoyar de manera remota y activa al tratamiento de pacientes de Alzheimer mayores de 65 años a través de la alerta hacia el médico y los familiares de anomalías en el comportamiento de estos.

Durante este artículo, primero, se detalla la metodología implementada para el desarrollo del proyecto para después hablar de los resultados encontrados.

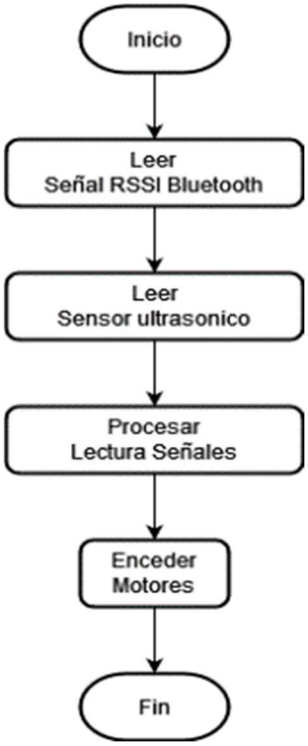


Fig. 3. Diagrama de bloques del circuito.

2. Metodología

Se desarrolló un modelo tridimensional virtual en SolidWorks para representar cómo se vería el robot en la vida real, de modo que se pueda decidir que materiales se van a emplear para la funcionalidad de este, materiales se refiere a qué plataformas, componentes mecánicos y electrónicos se piensan utilizar.

Se propone el uso de un chasis prefabricado para un robot diferencial. También se presenta filamentos de PLA para generar piezas que actúen como un apoyo para los sensores. Además, también se plantea implementar acrílico para construir una caja que sostenga la placa de control con sus respectivos circuitos para controlar los motores.

Diseñado el robot, se hizo un esquemático y simulación funcional en Proteus para definir cómo se relacionan los componentes electrónicos para el sistema de percepción de la solución, de modo que pueda registrar información del ambiente externo a ella para la navegación autónoma.

Finalmente, para probar la primera versión funcional se plantearon y evaluaron diferentes algoritmos en el IDLE de la plataforma Arduino, para que tomen la información del sistema de percepción para que el robot pueda permitirse una navegación autónoma y segura, mientras que, simultáneamente, sigue al paciente en tiempo real.

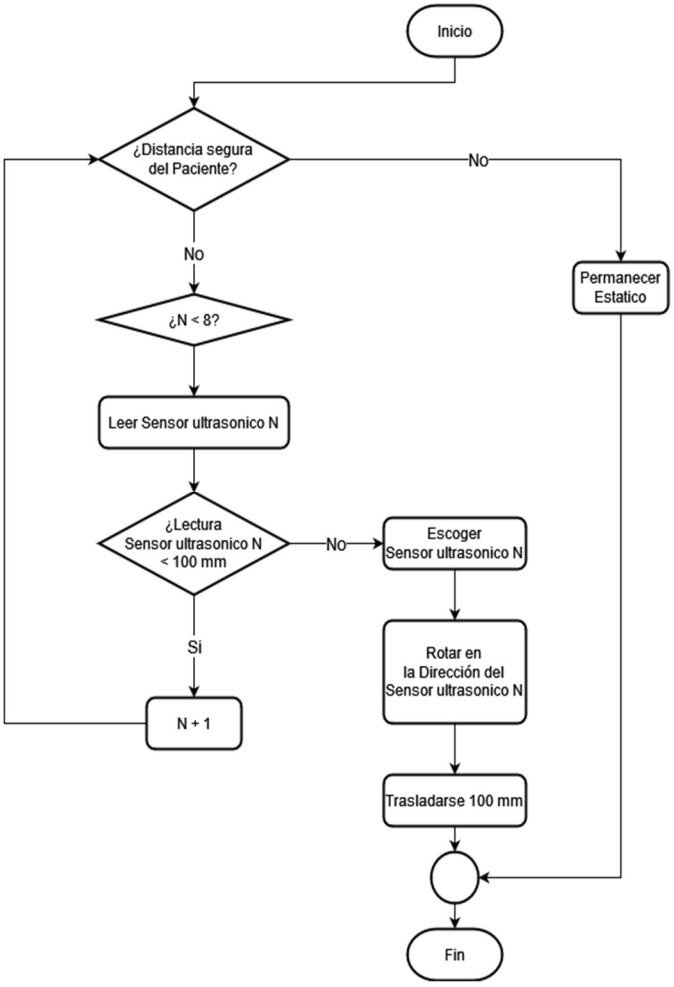


Fig. 4. Algoritmo de control.

3. Resultados

La solución propuesta a la problemática planteada consiste en un robot de asistencia social con la capacidad de monitorear posibles anomalías en el comportamiento mientras sigue al paciente en tiempo real. Para el seguimiento, el paciente hará uso de un rastreador (que consiste en un brazalete).

En este artículo, se expresarán lo resultados obtenidos hasta ahora con respecto al seguimiento de personas.

En esta primera propuesta, el sistema de percepción consiste en ocho sensores ultrasónicos que se encargan de detectar obstáculos cercanos al robot, pues toman

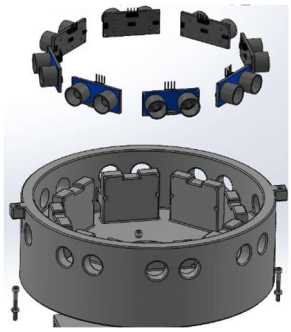


Fig. 5. Sensores ultrasónicos de DETER.

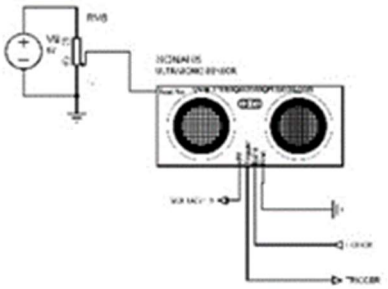


Fig. 6. Los ocho sensores ultrasónicos determinan que el robot tiene ocho direcciones hacia las cuales avanzar.

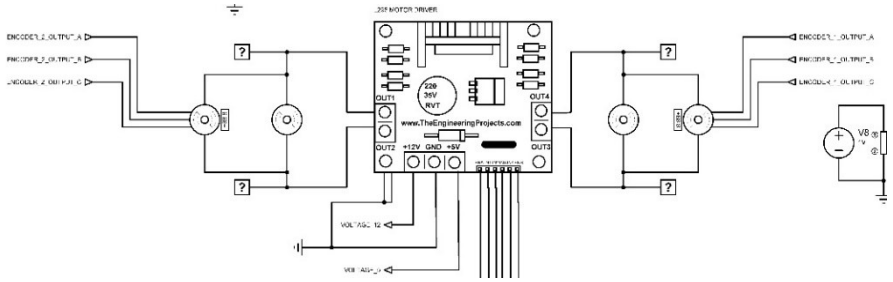


Fig. 7. El puente H actúa como un guía para los motores, diciéndoles la velocidad y el sentido a los mimos.

distancias de los objetos en las direcciones que ellos pueden cubrir. Las salidas de estos sensores son conmutadas por el multiplexor 74151, con el fin de no tener que emplear demasiados puertos de la placa.

Todas las señales de los ultrasónicos son mandadas hacia un Arduino UNO, que actúa como la placa encargada del procesamiento de la información dada por los sensores, así como el controlador para los actuadores.

El robot tiene ocho direcciones disponibles hacia las cuales avanzar porque tiene ocho sensores ultrasónicos instalados. Gracias a ellos, el analiza cuál de las direcciones se encuentran ocupadas y escoge la primera dirección que se encuentra disponible, entonces, procede a avanzar cierta distancia hacia ella.

Una vez se encuentra en dicha distancia, DETER evalúa si cumple con estar a la distancia necesaria del paciente. Si no lo está, vuelve a moverse en la primera dirección donde no haya obstáculos hasta que alcance al paciente.

Para moverse, el Arduino UNO les indica a los motores que se activen para girar y luego para moverse hacia la dirección especificada. Los motores son comandados por el driver L298, que es un puente H (circuito utilizado para controlar motores). Además, con el fin de medir y verificar la velocidad de ambos, los motores son medidos por unos encoders.

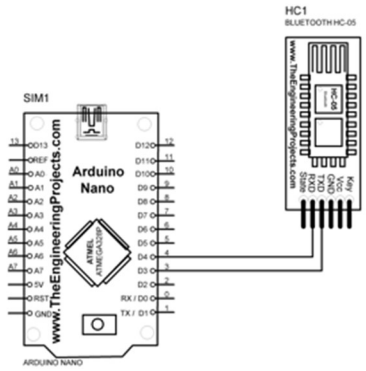


Fig. 8. Circuito del brazalete.

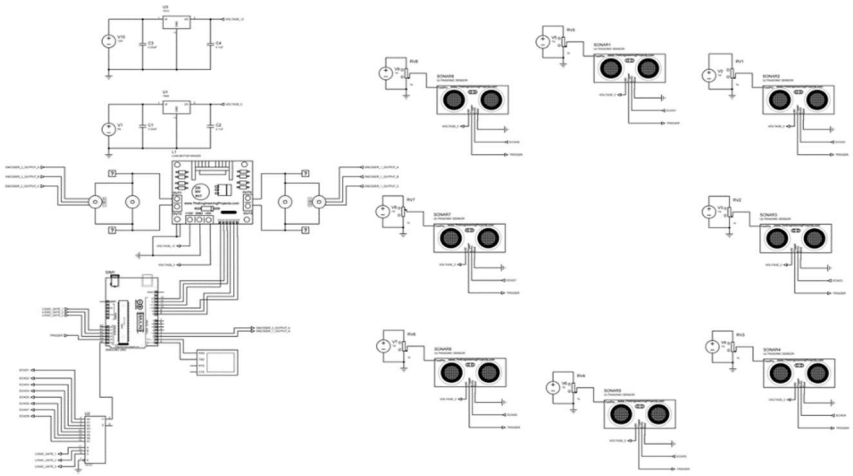


Fig. 9. El circuito final conecta los sensores ultrasónicos, el módulo bluetooth y los motores.

Para identificar y rastrear al paciente, el lleva puesto un brazalete que contiene un arduino NANO y un módulo bluetooth HC-05. El robot utiliza el RSSI (Received Signal Strength Indicator) de la conexión por bluetooth entre el Arduino del robot y el Arduino del brazalete que lleva el paciente.

En la Fig. 8 y 9, se presenta el circuito desarrollado hasta ahora para DETER:

4. Discusión y conclusiones

En la literatura, se ha visto a la robótica como una manera de suplir la necesidad de registrar y alertar de las anomalías en el comportamiento de pacientes de la enfermedad

del Alzheimer. En este artículo, se presenta un prototipo que cubre el primer intento de acercarse hacia dicha solución.

Para obtener mejores resultados en el futuro, se recomienda implementar un sistema de navegación más robusto y eficaz. Para ello, se recomienda instalar sensores adicionales con el fin de que la percepción del robot sea más precisa.

Así mismo, se propone implementar un algoritmo más robusto para el control de motores, pues el algoritmo actual es ineficaz para su uso en tiempo real dado que se demora mucho en seguir al paciente.

Referencias

1. Abubakar, M. B., Sanusi, K. O., Ugusman, A., Mohamed, W., Kamal, H., Ibrahim, N. H., Khoo, C. S., Kumar, J.: Alzheimer's disease: an update and insights into pathophysiology. Frontiers in Aging Neuroscience, vol. 14 (2022) doi: 10.3389/fnagi.2022.742408
2. Miller, R. K., Washington, K.: Chapter 105: Alzheimer's disease and dementia. Healthcare Business Market Research Handbook, 21st Edition, pp. 433–439 (2013)
3. Fathi, S., Ahmadi, M., Dehnad, A.: Early diagnosis of alzheimer's disease based on deep learning: A systematic review. Computers in Biology and Medicine, vol. 146, pp. 105634 (2022) doi: 10.1016/j.compbiomed.2022.105634
4. Napoli, C. D., Ercolano, G., Rossi, S.: Personalized home-care support for the elderly: A field experience with a social robot at home. User Modeling and User-Adapted Interaction, vol. 33, no. 2, pp. 405–440 (2022) doi: 10.1007/s11257-022-09333-y
5. Yuan, F., Anderson, J. G., Wyatt, T. H., Lopez, R. P., Crane, M., Montgomery, A., Zhao, X.: Assessing the acceptability of a humanoid robot for alzheimer's disease and related dementia care using an online survey. International Journal of Social Robotics, vol. 14, no. 5, pp. 1223–1237 (2022) doi: 10.1007/s12369-021-00862-x
6. Olde-Keizer-Richelle, A. C. M., Velsen, L., Moncharmont, M., Riche, B., Ammour, N., Signore, S., Zia, G., Hermens, H., N'Dja, A.: Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. Health and Technology, vol. 9, no. 4, pp. 595–605 (2019) doi: 10.1007/s12553-019-00320-9
7. Wang, L., Wu, J., Li, X., Wu, Z., Zhu, L.: Longitudinal control for person-following robots. Journal of Intelligent and Connected Vehicles, vol. 5, no. 2, pp. 88–98 (2022) doi: 10.1108/jicv-01-2022-0003
8. Islam, M. J., Hong, J., Sattar, J.: Person-following by autonomous robots: A categorical overview. The International Journal of Robotics Research, vol. 38, no. 14, pp. 1581–1618 (2019) doi: 10.1177/0278364919881683

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación