Advances in Artificial Intelligence

Research in Computing Science

Research in Computing Science, Año 19, Volumen 149, No. 5, mayo de 2020, 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 30 de agosto de 2019.

Research in Computing Science, year 19, Volume 149, No. 5, May 2020, is published monthly by the Center for Computing Research of IPN.

Volume 149(5)

Advances in Artificial Intelligence

Cuauhtémoc Sánchez-Ramírez
Giner Alor-Hernández
Jorge Luis García-Alcaraz
Alfonso Rojas Domínguez
Matías Alvarado (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2020

ISSN: 1870-4069

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

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

Thematic section: Intelligent Decision Support Systems for Industry Application

Regular Papers

Thematic Section
“Intelligent Decision Support Systems for Industry Application”

Cuauhtémoc Sánchez-Ramírez
Giner Alor-Hernández
Jorge Luis García-Alcaraz (eds.)

ISSN 1870-4069

Editorial for Thematic Section "Decision Support Systems for Industry Application"

Decision Support Systems (DSS) is integrated of models and methodologies to improve the industrial processes for this reason. In this volume, there are ten papers are presented that were carefully selected of 14 submissions about the use of different techniques for designing and developing Decision Support System (DSS) in industrial contexts. The papers were evaluated by an editorial board integrated for reviewers with international prestige in the area.

The papers were selected by considering the originality, scientific contribution to the field and technical quality of the papers. The articles were considered in the following industrial contexts: (1) Benefits of Green production; (2) Risk in storage and land transport of blood products; (3) Manufacturing technology to responsive supply chain strategy; (4) Sentiment analysis approach for drug; (5) Architecture for and IoT-based telecare system for the Elderly using Big Data Analytics; (6) Process for automatic generation of medical mobile applications using voice recognition; (7) life cycle cost analysis in wind energy projects; (8) Multi-agent systems for the inventory and routing assignment; (9) Multi-objective product allocation model in warehouses; (10) Convolutional neural network in Pseudo-distributed environment for classification of chest X-Ray images of patients with Pneumonia.

The editors would like to express their gratitude to the reviewers who kindly contributed to the evaluation of papers at all stages of the editing process. They equally thank the Editor-in-Chief, Prof. Grigori Sidorov, for the opportunity offered to edit this special issue and for providing his valuable comments to improve the selection of research works. Guest editors are grateful to the National Technological of Mexico for supporting this work and the National Council of Science and Technology (CONACYT) as part of the project named Thematic Network in Industrial Process Optimization.

Cuauhtémoc Sánchez-Ramírez (Instituto Tecnológico de Orizaba, Mexico)
Giner Alor-Hernández (Instituto Tecnológico de Orizaba, Mexico)
Jorge Luis García-Alcaraz (Universidad Autónoma de Ciudad Juárez, Mexico)
Guest Editors

January 2020

ISSN 1870-4069

Relationship among Green Production Benefits: A Causal Model

José Roberto Mendoza Fong[1], José Roberto Díaz Reza[1], Viridiana Reyes Uribe[2], Adrián Salvador Morales García[3], Jorge Luis García Alcaraz[3]

[1] Universidad Autónoma de Ciudad Juárez,
Departamento de Ingeniería Eléctrica y Computación,
Mexico

[2] Instituto Tecnológico de Ciudad Juárez,
Departamento de Ingeniería Industrial y Logística,
Mexico

[3] Universidad Autónoma de Ciudad Juárez,
Departamento de Ingeniería Industrial y Manufactura,
Mexico

{al164438, al164440, al194561}@alumnos.uacj.mx,
vreyes@itcj.edu.mx, jorge.garcia@uacj.mx

Abstract. Green production processes are becoming increasingly important, as consumers not only demand quality, economical and durable products, but also products that are environmentally friendly. Manufacturers are therefore wondering how feasible it is to transform their traditional production process into a green production process, since such transformation involves economic investment and the relationship between these two dimensions is unknown. This article presents a model of structural equations where four different types of benefits are associated: process benefits, quality benefits, market benefits, and green benefits that can be obtained by implementing a green production process and facilitating manufacturers' decision making. The model is validated with information from 559 responses from managers who have applied the concepts of green production processes in the Mexican manufacturing industry. The partial least squares technique is used to validate the models and the results indicate that the four benefits have a direct and positive effect on each other and the most significant is that there are process benefits and quality benefits.

Keywords: green production processes, environmental benefits, manufacturing industry, green product.

1 Introduction

Some of the most important factors in green activities are green production processes (GPP), which are important in the development of green supply chain management, as

the generation of green products will increase competitive advantages and improve the environment [1].

GPPs are a key step in achieving sustainability and ensuring the environmental, social, and economic aspects of manufacturing a product [2]. A GPP can be defined as a new manufacturing paradigm that incorporates diverse ecological strategies, drivers, and techniques to be more eco-efficient [3]. GPP refers to making products that consume less materials and energy, incorporating renewable and non-toxic materials, and reducing unwanted outflows, waste, emissions, and recycling [4].

However, implementing a GPP is not fast and manufacturers are wondering whether integrating green thinking into their production processes actually brings economic and environmental benefits. For this reason, this research presents a model of structural equations composed of four latent variables that associate a series of benefits that can be obtained by implementing a GPP in a certain organization.

1.1 Literature Review: Definition of Variables and Hypotheses

Manufacturers must design, manufacture, and distribute eco-friendly products to meet the demands of environmentally committed consumers. However, the question is whether manufacturers with a green perspective can realize benefits associated with the production process, sales, and quality. Fortunately, some researchers have shown that a GPP generates benefits [5-7].

First of all, one could ask what benefits can be gained in the production processes by having a GPP. For example, Gao, Xiao [8], Zhu and He [9] mention that GPP makes better use of resources, eliminates waste, achieves green process design and greater efficiency, competitiveness, productivity, and reduced cycle time.

Additionally, Shankar, Kumar [5] and Xie, Huo [10] mention that, by increasing efficiency, productivity, and reduction of cycle time, the quality of the product and process are improved and green processes and products are obtained. Therefore, the following hypothesis is defined:

> H_1: The *Process Benefits* have a direct and positive effect on the *Quality Benefits* in a GPP.

The efficiency of green production processes and the use of resources have a direct impact on customer service and the final consumer [11]. It is also believed that when designing sustainable products there will be an expansion in the market with consumers committed to the environment. [12] The following hypothesis is therefore defined:

> H_2: *Process Benefits* have a direct and positive effect on the *Market Benefits* in a GPP.

Therefore, manufacturers seek to distribute and develop eco-friendly products for national and international markets [13], which forces them to manufacture quality green products to improve their sales and increase their reputation with their customers.

Therefore, the implementation of a GPP improves the ecological performance of products and rebuilds an industrial system that reduces reprocesses and reduces cycle times for the customer. [14] Therefore, the hypothesis is defined:

H_3: *Quality Benefits* have a direct and positive effect on the *Market Benefits* in a GPP.

With GPP-enhanced production processes, all that remains is to analyze how consumers judge green products and how they influence the purchasing decision process. [15] In addition, consumers not only analyze the green products they buy, but also how they were manufactured, and the resources used, so that the brand image is related to what consumers think and the products they buy, which defines the following working hypothesis:

H_4: The *Process Benefits* have a direct and positive effect on the *Green Benefits* in a GPP.

When a brand is perceived as having a green image, its products are linked with quality in the mind of the consumer. A green product image helps companies attract more customers by affecting consumer choice and improving consumer brand loyalty. [10] Since there are many consumers who want to buy products from companies that respect the environment and few companies that generate these types of products, the consumption of these types of products has increased. [16] Therefore, the following working hypothesis is defined:

H_5: *Quality Benefits* have a direct and positive effect on the *Green Benefits* in a GPP.

On the other hand, investing in GPP innovation helps prevent companies from facing environmental protests and legal sanctions, allows them to develop new market opportunities, and improves customer service. [17] Now, companies have an ecological competition to increase sales, strengthen the ecological image, and improve their acceptance in society. [10] The following hypothesis is therefore proposed:

H_6: *Market Benefits* have a direct and positive effect on *Green Benefits* in a GPP.

The hypotheses defined above are illustrated in Fig. 1.

2 Methodology

2.1 Collection of Information

A literature review is carried out in relation to the GSC and GPP, approximately 100 different scientific articles were analyzed to identify the most mentioned and commonly obtained benefits of applying GPP, which were classified into categories (as indicated in Table 1) and are the items in each latent variable of the model analyzed.

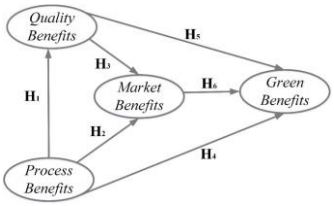


Fig. 1. Hypothesis.

Table 1. Benefits Classification.

Process Benefits [10, 18, 19]	*Manufacturing Benefits* [20-22]
Reduction of cycle time. Improved product design. Reduction of cycle time. Higher competitivity, productivity and efficiency. Better use of available resources.	Expansion of the world market. Better customer service. Improved reputation with customers and competitors.
Quality benefits [3, 23, 24]	*Green benefits* [13, 25, 26]
Improved process quality. Less product reprocessing. Higher quality in the final product.	Image of a sustainable company. Better acceptance of products in society. Integration of the company in society.

Table 2. Descriptive analysis of latent variables and items.

Process Benefits	Median	IR
Reduction of cycle time	3.48	1.808
Improved product design	3.718	1.577
Higher competitivity, productivity and efficiency	3.721	1.514
Better use of available resources	3.743	1.52
Manufacturing Benefits		
Increasing the quality of your processes	3.697	1.549
Reduced product reprocessing	3.604	1.626
Increase in the quality of the final product	3.786	1.563
Market Benefits		
Expansion of the world market	3.695	1.66
Better customer service	3.805	1.547
Improved reputation with customers and competitors	3.823	1.587
Green Benefits		
Image of a sustainable company	3.56	1.704
Better acceptance of products by society	3.693	1.66
Integration of the company into society	3.793	1.683

With these benefits, a survey was designed as a tool to gather information on the situation of manufacturing companies, which is applied to managers and engineers responsible for GGP and which was answered on a Likert scale. The information obtained is captured in a database built in the Statistical Software SPSS 24 ®.

2.2 Items Analysis

The median is used as a measure of central tendency. Low values indicate that the benefit is not obtained, while high values indicate that they are always obtained. In addition, the interquartile range (IR) is used as a measure of dispersion, where low values indicate consensus among responders and high values indicate lack of consensus.

2.3 Execution of the Model of Structural Equations

The structural equation model (SEM) is evaluated in the software WarpPls 6.0 ®, which integrates the partial least squares technique that uses standardized values and is used in small samples and data that do not tend to normality.

The following indices are used to validate the latent variables in Figure 1: Average R-square (ARS), Average adjusted R-square (AARS), Average path coefficient (APC), Average variation inflation factor (AVIF), Average total collinearity VIF (AFVIF) and Tenenhaus Index.

For APC, ARS y AARS, p-values are analyzed, setting 0.05 as the limit and testing null hypotheses in which APC, ARS y AARS = 0, against the alternative hypothesis APC, ARS y AARS $\neq 0$. The values of AVIF and AFVIF should be less than 5 and for the Tenenhaus Index (GoF), it is recommended to have values greater than 0.36.

In the SEM, three types of effects are measured in the model: direct, indirect and total. The direct effects are the arrows that directly connect two latent variables, the indirect effects are represented by routes with two or three segments and the total effects is the sum of the direct and indirect effects. For statistical significance a 95% confidence level is used, testing the null hypothesis: $\beta i = 0$, versus the alternative hypothesis: $\beta i \neq 0$.

3 Results

3.1 Descriptive Item Analysis

Table 2 shows the descriptive analysis of the items, where the second column indicates the median and it is observed that 12 of 13 thirteen benefits have an average greater than 3,500 which denotes that these benefits are important for a GPP. The last column illustrates the IR and it can be seen that all benefits have a value less than 2 and it is concluded that there is consensus among respondents regarding the importance of these benefits.

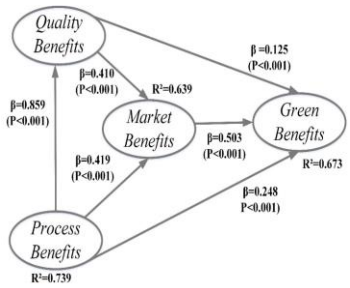


Fig. 2. Evaluated model.

Table 3. Validation of hypotheses.

Hypothesis	Independent Variable	Dependent Variable	β	Effect Size	P-Value	Decision
H_1	*Process Benefits*	*Quality Benefits*	0.859	0.739	P< 0.001	Accepted
H_2	*Process Benefits*	*Market Benefits*	0.419	0.323	P< 0.001	Accepted
H_3	*Quality Benefits*	*Market Benefits*	0.410	0.316	P< 0.001	Accepted
H_4	*Process Benefits*	*Green Benefits*	0.248	0.184	P< 0.001	Accepted
H_5	*Quality Benefits*	*Green Benefits*	0.125	0.091	P< 0.001	Accepted
H_6	*Market Benefits*	*Green Benefits*	0.503	0.398	P< 0.001	Accepted

3.2 Equation Model Validation

Fig. 2 shows the model obtained, where each arrow indicates a direct effect between two latent variables, and this includes a beta parameter. (β), a p-value for the hypothesis test and an R2 value as a percentage of the explained variance of the latent variables. The validation indices of the latent variables are:

- Average path coefficient (APC) = 0.427, P<0.001,
- Average R-squared (ARS) = 0.684, P<0.001,
- Average adjusted R-squared (AARS) = 0.683, P<0.001,
- Average block VIF (AVIF) =3.854, acceptable if <= 5,
- Average full collinearity VIF (AFVIF) = 3.813, acceptable if <= 5,

Table 4. Sum of indirect and total effects.

Sum of indirect effects			
		To	
From	*Quality Benefits*	*Market Benefits*	*Green Benefits*
Process Benefits		0.353 (P<0.001) ES=0.272	0.495 (P<0.001) ES=0.369
Quality Benefits			0.206 (P<0.001) ES=0.150
Sum of total effects			
Process Benefits	0.859 (P<0.001) ES=0.739	0.771 (P<0.001) ES=0.595	0. 743 (P<0.001) ES=0.554
Quality Benefits		0.410 (P<0.001) ES=0.316	0.331 (P<0.001) ES=0.241
Market Benefits			0.503 (P<0.001) ES=0.398

— Tenenhaus GoF (GoF) = 0.728, small >= 0.1, medium >= 0.25, large >= 0.36.

Table 4 presents the validation of the six defined hypotheses and the conclusions.

3.3 Total and Indirect Effects

At this point, Table 5 will be presented, which contains the indirect effects as well as the total effects presented by the model. First, you will see the indirect effects, which are in two and three segments. Then the total effects are presented, which is the sum of direct effects (β) and indirect effects.

4. Conclusions

It is concluded that by transforming a traditional production process to a GPP, benefits are obtained in processes, quality, market and image; since the model proposed here demonstrates quantitatively that the four latent variables have a direct, positive and significant effect on each other.

It is observed that when obtaining *Process Benefits* the obtaining of the *Quality Benefits* is guaranteed, since this is the biggest and significant effect in all the model and which was of 0.859 (H_1), in addition to the fact that Das, Rukhsana [3, 7] demonstrated that when it comes to optimizing product processes, these improvements will be seen in the quality of the product, as well as in the optimization of the use and consumption of resources.

Finally, today, apart from the fact that the product that the consumer is looking to be green, it is important the image of the brand of the company that is manufacturing the product is green.

This is also important when marketing the product because consumers are focusing more and more on how that green product is produced and this is verified in the $H_{6,}$ since the market benefits are fundamental to be able to obtain the *green benefits*, a statement that detonates [15] in our investigation.

References

1. Watkins, L., Aitken, R., Mather, D.: Conscientious consumers: A relationship between moral foundations, political orientation and sustainable consumption. Journal of Cleaner Production, 82, pp. 137–146 (2016)
2. Bonvoisin, J., Stark, R., Seliger, G.: Field of research in sustainable manufacturing, in sustainable manufacturing: Challenges, solutions and implementation perspectives. Stark, R., Seliger, G., Bonvoisin, J. Editors. Springer International Publishing, pp. 3–20 (2017)
3. Das, A., Rukhsana, Chatterjee, P.: Green Manufacturing: Progress and Future Prospect, in Reference Module in Materials Science and Materials Engineering. Elsevier (2019)
4. Alayón, C., Säfsten, K., Johansson, G.: Conceptual sustainable production principles in practice: Do they reflect what companies do?. Journal of Cleaner Production, 141, pp. 693–701 (2017)
5. Shankar, K.M., Kumar, P.U., Kannan, D.: Analyzing the drivers of advanced sustainable manufacturing system using AHP approach. Sustainability, 8(8) pp. 824 (2016)
6. Aboelmaged, M.: The drivers of sustainable manufacturing practices in egyptian SMEs and their impact on competitive capabilities: A PLS-SEM model. Journal of Cleaner Production, 175, pp. 207–221 (2018)
7. Mendoza-Fong, J.R., García-Alcaraz, J.L., Díaz-Reza, J.R., Jiménez, E.: The role of green attributes in production processes as well as their impact on operational, commercial, and economic benefits. Sustainability, 11(5), pp. 1294 (2019)
8. Gao, J., Xiao, Z., Wei, H., Zhou, G.: Active or passive? Sustainable manufacturing in the direct-channel green supply chain: A perspective of two types of green product designs. Transportation Research Part D: Transport and Environment, 65, pp. 332–354 (2018)
9. Zhu, W., He, Y.: Green product design in supply chains under competition. European Journal of Operational Research, 258(1), pp. 165–180 (2017)
10. Xie, X., Huo, J., Zou, H.: Green process innovation, green product innovation, and corporate financial performance: A content analysis method. Journal of Business Research, 101, pp. 697–706 (2019)
11. Seth, D., Rehman, M.A.A., Shrivastava, R.L.: Green manufacturing drivers and their relationships for small and medium (SME) and large industries. Journal of Cleaner Production, 198, pp. 1381–1405 (2018)
12. Zhu, Q., Sarkis, J.: Green marketing and consumerism as social change in China: Analyzing the literature. International Journal of Production Economics, 181, pp. 289–302 (2016)
13. Abu-Seman, N.A., Govindan, K., Mardani. A., Zakuan, N., Zameri-Mat Saman, M., Hooker, R.E., Ozkul, S.: The mediating effect of green innovation on the relationship between green supply chain management and environmental performance. Journal of Cleaner Production, 229, pp. 115–127 (2019)

14. Zhang, X., Ming, X., Liu, Z., Qu, Y., Yin, D.: General reference model and overall frameworks for green manufacturing. Journal of Cleaner Production, 237, pp. 117757 (2019)
15. Yi Chang Yang: Consumer behavior towards green products. Journal of Economics, Business and Management, 5(4), pp. 160–167 (2017)
16. Xie, X., Huo, J., Qi, G., Xiaoguo-Zhu, K.: Green process innovation and financial performance in emerging economies: Moderating effects of absorptive capacity and green subsidies. IEEE Transactions on Engineering Management, 63(1), pp. 101–112 (2016)
17. Dai, R., Zhang, J.: Green process innovation and differentiated pricing strategies with environmental concerns of South-North markets. Transportation Research Part E: Logistics and Transportation Review, 98, pp. 132–150 (2017)
18. Singh, A., Philip, D., Ramkumar, J., Das, M.: A simulation based approach to realize green factory from unit green manufacturing processes. Journal of Cleaner Production, 182, pp. 67–81 (2018)
19. Jamali, M.-B., Rasti-Barzoki, M.: A game theoretic approach for green and non-green product pricing in chain-to-chain competitive sustainable and regular dual-channel supply chains. Journal of Cleaner Production, 170, pp. 1029–1043 (2018)
20. Sharma, V.K., Chandna, P., Bhardwaj, A.: Green supply chain management related performance indicators in agro industry: A review. Journal of Cleaner Production, 141, pp. 1194–1208 (2017)
21. Madani, S.R., Rasti-Barzoki, M.: Sustainable supply chain management with pricing, greening and governmental tariffs determining strategies: A game-theoretic approach. Computers and Industrial Engineering, 105, pp. 287–298 (2017)
22. Ji, J., Zhang, Z., Yang, L.: Carbon emission reduction decisions in the retail -dual- channel supply chain with consumers' preference. Journal of Cleaner Production, 141, pp. 852–867 (2017)
23. Seth, D., Seth, N., Dhariwal, P.: Application of value stream mapping (VSM) for lean and cycle time reduction in complex production environments: A case study. Production Planning & Control, 28(5), pp. 398–419 (2017)
24. Dangelico, R.M., Pujari, D., Pontrandolfo, P.: Green product innovation in manufacturing firms: A sustainability-oriented dynamic capability perspective. Business Strategy and the Environment, 26(4), pp. 490–506 (2017)
25. Hong, Z., Guo, X.: Green product supply chain contracts considering environmental responsibilities. Omega, 83, pp. 155–166 (2019)
26. Xiao, Y., Yang, S., Zhang, L., Kuo, Y.H.: Supply chain cooperation with price-sensitive demand and environmental impacts. Sustainability, 8(8), pp. 716 (2016)

ISSN 1870-4069

Operational Risk in Storage and Land Transport of Blood Products

Juan Carlos Osorio Gómez, Mayerli Daniela Naranjo Sánchez,
Nataly Agudelo Ibarguen

Escuela de Ingeniería Industrial,
Universidad del Valle,
Colombia

{juan.osorio, mayerli.naranjo,
nataly.agudelo}@correounivalle.edu.co

Abstract. Blood transfusions contribute to saving and improving the quality of life of thousands of people in the world every day. There is no alternative to satisfy the demand for blood in medical procedures and there is no substitute for human blood (blood is supplied voluntarily by donors). Therefore, blood should be seen as a scarce resource in the world. Currently, the donation rate in Colombia is considered low. However, the donor deficit is not the only problem for this type of institution, other factors within the biological and logistical control of the chain directly affect the safety and availability of blood, making the functioning of the blood supply chain an interesting study problem to treat. This supply is exposed to the possibility of unexpected events affecting the normal functioning of its activities, this is called operational risk. These risks affect the performance of organizations and it is therefore important to work on their identification and prioritization so that organizations can guide their efforts to mitigate or eliminate them. Due to the social and economic importance of improving blood supply systems, the present work identifies, Prioritizes and provides mitigation actions for the impact of operational risks associated with the storage and land transport of blood products in Colombia.

Keywords: supply chain risk management, operational risk, fuzzy-QFD methodology, blood banks, probability-impact matrix.

1 Introduction

According to the World Health Organization (WHO), blood donations help save lives and improve people's health [1]. Hence, the different authorities on the issue see the need to launch awareness campaigns on the importance of blood donation.

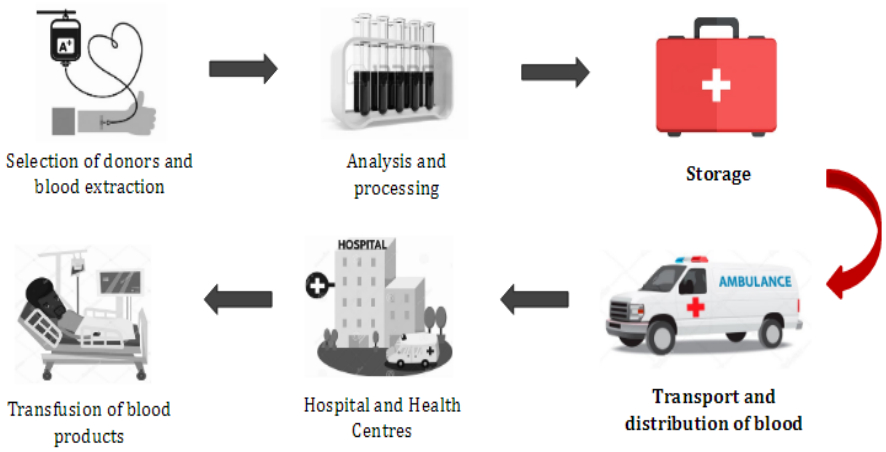


Fig. 1. Blood supply chain.

A sample of the work carried out by these entities is that, globally, approximately 112,5 million units of blood are donated per year, according to data collected and published by WHO as of June 2017. Moreover, in Colombia, only between 2014 and 2015, according to figures revealed by the National Institute of Health, around 792.000 units of blood were collected per year [2].

However, the blood that is donated is not sufficient to meet the demand for this supply. In Colombia, the rate of donations is 15,4 per thousand inhabitants when ideally there would be between 30 and 40 donors for every thousand people in the country, according to Gabriel Cubillos, member of the board of directors of the Colombian Association of Blood Banks and Transfusion Medicine ACOBASMET (2016). In the annual report of the Blood Bank Network for 2017, it is stated that:

> *"Considering on average in Colombia a unit of red blood cells has an economic value for the system of $291.733 [3] and more than 30,000 units of red blood cells were discarded from blood banks for controllable causes, it is estimated that the Health System lost in processing costs about 8,700 million pesos ($8'751.900.000), which can result in a detriment for blood banks, and failures related to demand satisfaction, which for this component was 88,8% and 90% for platelets".*

Incorrect risk management in this process not only entails a huge loss of economic resources for the country and the health sector, but directly affects the patient's health and quality of life [4], the latter being the most important impact. This is why risk management takes on prominence, seeking to reduce, eliminate or avoid disturbances of the physical and information flows that affect the interaction of the different links in the supply chain [5]. The chain can be presented graphically in Figure 1.

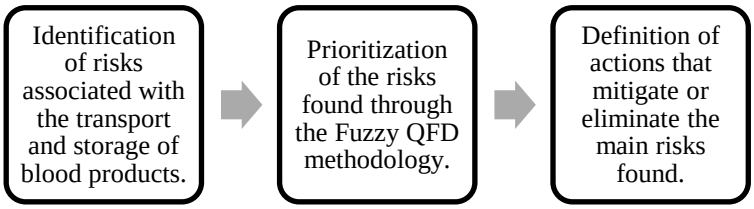


Fig. 2. Methodological design.

Table 1. Linguistic scale.

Linguistic Scale	**Very low (VL) Insignificant (I)**	**Low (L) Minor (Mi)**	**Medium (M) Moderate**	**High (H) Critical (C)**	**Very high (VH) Severe (S)**
Numerical equivalence	1	2	3	4	5

2 Methodology

The methodological design for the development of the project is presented below in Figure 2. The methodology will have three phases.

2.1 Identification of Operational Risks Associated with the Transport and Storage of Blood Products

From the literature and through the application of questionnaires to experts, the main risks associated with the transport and storage of blood products are identified. The application of the questionnaire was carried out individually and allowed the experts to rate the risk, both in probability and impact, using the linguistic scale illustrated in Table 1.

After the application of the questionnaire, responses are consolidated and processed according to the scale. The data obtained apply Equation 1 and Equation 2 to obtain the percentages of risk application and weighted averages of probability of occurrence and magnitude of impact:

Equation 1: Weighted average of the magnitude of risk i:

$$\bar{X}_i = \frac{\sum_{j=1}^{n}(B_{i,j} \times M_{i,j})}{n} \quad ; \quad \forall\, i. \tag{1}$$

Equation 2: Weighted average probability of risk i:

$$\bar{Y}_i = \frac{\sum_{j=1}^{n}(B_{i,j} \times P_{i,j})}{n} \quad ; \quad \forall\, i. \tag{2}$$

$\bar{X}_i$ =Weighted average of the magnitude of risk i,

$\overline{Y}_i$ = Weighted average probability of risk i,
$B_{i,j}$ = Expert's criterion j if i applicable as risk (1,0),
$M_{i,j}$ = Expert's qualification j on the impact of risk i,
$P_{i,j}$ = Expert's qualification j on the probability of risk i.

Once these data are obtained, the Impact Matrix is established. Risks in the green section are negligible risks, risks in the yellow section are risks with a moderate probability of occurrence, while the risks of the orange and red sections are critical risks and it is those that are selected for the next phase.

2.2 Prioritization of Operational Risks Through Fuzzy-QFD

The following methodology is based on the article by [6], and also the specific methodological proposal for prioritization described in [7] in which the following phases are determined:

- Identify the internal variables "What's",
- Determine the relative importance of "What's",
- Identify strategic objectives or "How's",
- Determine the correlation between "What's and How's",
- Determine the weight of the "How's",
- Determine the impact of risk on strategic objectives "How's",
- Prioritize the risks involved.

2.3 Operational Risk Mitigation Strategies or Actions

On the basis of the above qualifications, strategies must be defined in order to mitigate or eliminate the risks present in the process and thereby improve the process. As a final step it is important to emphasize in the implementation of actions aimed at transferring, eliminating and/or reducing the risks of the process and to apply focused strategies on the individual or associated machinery [8].

3 Results

The results are then presented using the methodology set out in the previous chapter for the identification and prioritization of operational risks associated with the transport and storage of blood products in Colombia, with the aim of developing actions aimed at mitigation and risk reduction.

Table 2. Validation and weighted averages of operational risks.

Risk Description	Id	Probability of Occurrence	Impact
Equipments that make up the cold chain without meeting minimum specifications and international standards (WHO).	R_1	2,74	4,05
Equipment used in unstandardized transport and storage.	R_2	2,84	3,95
Lack of availability of technical support, spare parts and maintenance services for cold chain equipment.	R_3	2,53	3,79
Cuts in the power supply that prevent the continuous operation of cooling equipment.	R_4	2,05	4,11
Use of picnic fridges due to the shortage of portable refrigerators suitable for blood conservation.	R_5	2,63	2,63
Equipment without sufficient technology for temperature control (Devices with built in temperature control failures and maintenance threshold overruns alarms).	R_6	2,95	4,05
Equipment's inability to maintain a stable temperature under extreme ambient temperature and humidity conditions (from +2 °C to +6 °C, the operating temperature of the equipment being +4°C).	R_7	2,53	3,95
Lack of or inadequate advice from the operator/transporter to the sender on the technical specifications for the transport of the total blood, hemocomponents and samples.	R_8	2,47	4,00
Lack of professionals specialized in Transfusion Medicine.	R_9	2,79	3,37
Absence/Error in the labelling of transport units.	R_{10}	2,26	4,05
Use of FIFO policies (First components processed, first to be sent for use).	R_{11}	0,89	1,42
Failure in the absorbent material between the primary container (total blood unit or hemocomponent) and the secondary container (hermetically sealed plastic bag) in order to prevent leakage affecting the outer container (refrigerator where blood is transported).	R_{12}	1,74	2,89
Documentation relating to the dispatch and transport of incomplete or poorly registered total blood.	R_{13}	2,00	2,89
No standardised or digital recording system.	R_{14}	2,32	2,68
It's a traffic accident.	R_{15}	2,26	3,68
Inefficient route for the transport of total blood, blood cells and samples.	R_{16}	1,79	2,53
Technical failures in the vehicle.	R_{17}	1,84	3,00

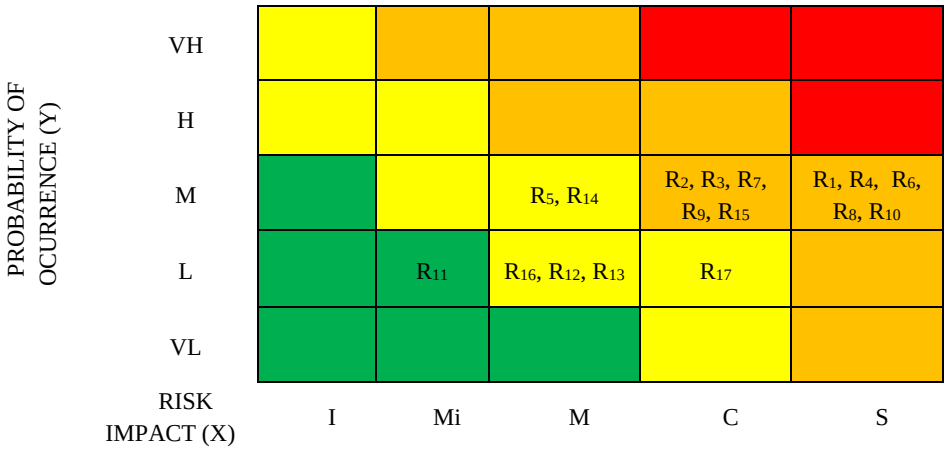


Fig. 3. Risk probability-impact matrix.

3.1 Identification of Operational Risks

Table 2 presents the identified risks and the results of the application of the questionnaire.

Once the risks are identified with their probability of occurrence and the impact they generate, these results are placed in a probability-impact matrix, taking into account the rating each expert considers for each risk. This is shown in figure 3.

3.2 Prioritization of Operational Risks

Thereafter, operational risks will be prioritized in the processes of transport and storage of blood products, with the FQFD methodology as mentioned above based on Chapter 5 of [9], and the risks to be taken into account for this prioritization will be the risks located in the red and orange areas of the probability matrix – impact of Figure 3. These are the risks of greater inference in the aforementioned processes.

The result of applying the aforementioned methodology is presented in Table 3, which shows that the critical risks are R6, R2, R10 and R7, which also have been classified as equipment and human hazards.

In accordance with this prioritization, action plans are established to help reduce or mitigate the impact of these plans in relation to the strategic objectives of the process, for this purpose the eight defined risks will be classified into two categories, taking into account that they are risks arising as a result of people and equipment.

3.3 Actions Aimed at Mitigation

Once the risk prioritization stage was completed, the choice was made to group the risks according to their origin, resulting in two groups: risks associated with equipment

and risks associated with staff. For each grouping, the Cause-Effect tool is used to investigate the causes related to each type of risk. Subsequent to the identification of cases, interviews with experts are conducted to identify actions to mitigate or eliminate prioritized risks.

After proper validation with the expert panel, the following actions were identified:

Actions Aimed at Risk Mitigation for Equipment:

- Appropriate packs must be available to maintain temperature over long distances. These packages are: Cardboard box, isothermal box and isothermal fridge. Additionally, for the protection of hemocomponents inside the refrigerator can be used damping materials such as absorbent paper or bubble paper among others [10].
- A preventive review program is recommended for all the teams that make up the chain, taking into account the specialty of each equipment and the times required in case of full or partial replacement and the availability of their spare parts.
- A program of revision and calibration must be in accordance with international standards and minimum specifications for the use of cold chain equipment for the process.

Actions Aimed at Risk Mitigation for Staff:

- Human resource needs should be identified and ensured, including operational and personnel for general storage and transport activities, which should be competent, technically updated and trained for the position held [11].
- Staff should be aware of existing standards and working arrangements, continuous training in the quality system and sufficient time to do the work and for activities such as inspection and verification[11].
- It should be borne in mind that individual roles and responsibilities are clearly defined, documented and disseminated in a way that avoids gaps and overlaps. The responsibilities assigned to each person will not be so numerous as to constitute a risk to the quality of the products [11].
- It is recommended to have written procedures on staff selection, training and training processes [11].

4 Conclusions

The results of the prioritization made it clear that the critical risks of the process are aimed at the teams and people involved throughout the chain. It is important for the country and public health to invest in blood banks by the state using prioritization techniques such as the one developed in this work to channel and target the best investment opportunities.

Table 3. Results of prioritization.

N°	Type of Risk	Description of the Risk	IPRF
		Very high risk	**552,0**
6	Risks of the equipment	Equipment without sufficient technology for temperature control (Devices with built in temperature control failures and maintenance threshold overruns alarms).	486,4
2	Risks of the equipment	Equipment used in unstandardized transport and storage.	470,3
N°	**Type of Risk**	**Description of the Risk**	**IPRF**
10	Risks of the Personnel	Absence/Error in the labelling of transport units.	442,0
7	Risks of the equipment	Equipment's inability to maintain a stable temperature under extreme ambient temperature and humidity conditions (from +2 ℃ to +6 ℃, the operating temperature of the equipment being +4 ℃).	439,4
		Risk High	**431,1**
8	Risks of the Personnel	Lack of or inadequate advice from the operator or transporter to the sender on the technical specifications for the transport of the total blood, hemocomponents and samples.	427,7
1	Risks of the equipment	Equipment's that make up the cold chain without meeting minimum specifications and international standards (WHO).	419,0
3	Risks of the equipment	Lack of availability of technical support, spare parts and maintenance services for cold chain equipment.	396,2
9	Risks of the Personnel	Lack of professionals specialized in Transfusion Medicine.	370,4

Energy options for the cold chain as solar powered refrigerators would set a long term precedent for handling the hostile climates of the Colombian geography, as these places are the most affected by the states of the equipment and energy problems.

The most important limitation is related to the questionnaire. It is sent to many experts, but it is not easy for them to respond.

For future work, continuing research on risk management around the blood supply chain, the issue of risk quantification could be addressed by considering also the impact

of the social component of blood supply. Future work could consider the economic elements related to these risks.

References

1. Organización Mundial de la salud: OMS: ¿Por qué es importante donar sangre?. Who, pp. 1 (2016)
2. Vengoechea, M.: Bancos de sangre en Colombia, en saldo rojo (2016)
3. Forero, M.I.B.: Informe anual red de sangre (2017)
4. Departamento de Tecnologías Sanitarias Esenciales Organización Mundial de la Salud: La cadena de frio de la sangre. Guía para la Adquisición de Equipos y Accesorios, OMS, pp. 77 (2004)
5. Giannakis, M., Louis, M.: A multi-agent based framework for supply chain risk management. Journal of Purchasing Supply Management, 17, pp. 23–31 (2011)
6. Osorio-Gómez, J. C.: Fuzzy QFD for multicriteria decision making - Application example. Prospectiva, 9(2), pp. 22–29 (2011)
7. Osorio-Gomez, J.C., Manotas-Duque, D.F., Rivera, L., Canales, I.: Operational risk prioritization in supply chain with 3PL using Fuzzy-QFD. In New Perspectives on Applied Industrial Tools And Techniques, Management An Industrial Engineering, pp. 91–109 (2018)
8. Lavastre, O., Gunasekaran, A., Spalanzani, A.: Supply chain risk management in French companies. Decision Support System, 52(4), pp. 828–838 (2012)
9. Osorio, J.C., Manotas, D.F., Rivera, L.: Priorización de Riesgos Operacionales para un Proveedor de Tercera Parte Logística - 3PL. Información Tecnologica, 28(4), pp. 135–144 (2017)
10. Instituto Nacional de Salud: Lineamiento para cadena de frio en transporte de hemocomponentes version 02 (2019)
11. Instituto Nacional de Cancerología: Ministerio de la Protección Social. 2006 (2007)

ISSN 1870-4069

Effect of AMT on Responsive Supply Chain Strategy, Pull System and Responsiveness to Market

José Roberto Díaz Reza[1], José Roberto Mendoza Fong[1],
Adrián Salvador Morales García[2], Jorge Luis García Alcaraz[2]

[1] Universidad Autónoma de Ciudad Juárez,
Departamento de Ingeniería Eléctrica y Computación,
Mexico

[2] Universidad Autónoma de Ciudad Juárez,
Departamento de Ingeniería Industrial
Mexico

{al164438, al164440, al194561}@alumnos.uacj.mx,
jorge.garcia@uacj.mx

Abstract. This paper presents a model of structural equations in which four variables (advanced manufacturing technology, Pull system, responsiveness to market and, responsive supply chain strategy) are related using six hypotheses. The objective of the research is to measure the effect that occurs between these variables and identify the most important activities that have the greatest effect on the others. The model is statistically validated with information of 254 responses to a questionnaire applied in the manufacturing industry and the partial least squares technique is used. The results indicate that advanced manufacturing technologies indirectly help companies to be able to respond to changes in demand and allow them to offer a rapid response in the changing market through the implementation of a pull system.

Keywords: AMT, pull system, supply chain, responsiveness supply chain.

1 Introduction

With complex and globalized production systems, supply chains (SC) perform essential functions in coordination with various commercial entities and in connecting offer with demand. [1]. In addition, with the growing proliferation of products and shorter technology life cycles, time to market is critical to avoid obsolete inventories [2] and therefore, manufacturing companies must continually update their product offerings, while remaining competitive [3] in that sense, managers must adapt the characteristics of the SC [3].

To achieve the above and maintain or improve the competitive position of their companies, managers must continue to improve their operations [4] and resort to the implementation of advanced manufacturing technology (AMT) in their SC, which must

be aligned with Product development decisions that must be designed at low cost and delivered in specific time and quality [3]. Thus, AMT facilitates the development of a successful long-term strategy of the SC with the client [5].

This strategy of the SC and AMT should be integrated into the production system and therefore, companies use pull systems, which support the administration of component inventories and their applications are varied. Fortunately, there are works in which these variables have been related; for example, Koufteros [6] has linked through a SEM the Pull system with preventive maintenance, where he analyzes the improvement of the setup as an independent variable and the reliability of delivery as a dependent variable. Also in Díaz-Reza, García-Alcaraz [7], AMTs have been related to the benefits that are obtained from their correct implementation within the production systems. The main problem is that these two tools have not been analyzed together and, in addition, have not been related to the strategy and speed of response of the supply chain. That is why, the aim of this research is to quantify, using a structural equation model, the effect of AMT on a pull system on the responsiveness of companies and their strategy in the supply chain within the industry of manufacturing. The rest of the article addresses the definitions for each of the variables, the methodology, the results and the conclusions.

2 Literature Review and Hypotheses

2.1 Advanced Manufacturing Technology (AMT)

Advanced manufacturing is understood as the use of modern technologies to deliver existing products and new products to the market, and also focuses on the improvement of design and manufacturing processes in all areas, along with the integration of information technology systems throughout the SC [8]. In manufacturing, technology is incorporated into products, in the physical processes by which they are manufactured and, increasingly, in the management systems that control all operations, which are traditionally known as AMT [9]. The benefits of AMT is that it helps improve cost, quality, flexibility and delivery times. [9]. AMT has operational, technical superiority and other intangible benefits, compared to traditional systems [10-12], such as: increased competitiveness, lower production cost, higher value for money, employment of fewer people, minimum inventories, product quality, flexibility, among others. In this investigation, AMT is evaluated by the following items, which are associated with the handling of materials and support the SC:

> *Automated parts loading/unloading, Automated guided vehicles, AGV´s and, Automated storage-retrieval systems, AS/RS.*

2.2 Responsive Supply Chain Strategy (RSCS)

The effective strategy SC refers to the process configuration of a supply network so that operation directly support corporate strategy. [13]. However, it is likely that the introduction of new products and services or the entry into new markets will be more

successful if it is accompanied by innovative CS designs, innovative practices and technology enablement [14]. In this investigation, the supply chain strategy is evaluated with the following items:

> *Wider product range, Offer new products more frequently and, Offer more innovative products.*

To achieve this, AMTs allow customers to absorb the options to create a new design and transform it into tangible products [15]. Therefore, the capabilities to achieve flexibility of products and processes contribute greatly to maintaining the competitive advantage among the companies that implement these technologies [16]. In that sense, the following hypothesis is established:

> *H_1: AMT has a direct and positive effect on responsive supply chain strategy.*

2.3 Pull System

In a pull system, work in process (WIP) is extracted through down-demand operation instead of the traditional push approach. The efficiency of the pull system is due to the visible signals and the WIP limit [17]. Pull systems are evaluated by the following items:

> *Undertaking actions to implement pull production (e.g., reducing batches, setup times, using kanban systems, etc.) and, Planned effort to implement pull production (e.g. reducing batches, setup times, using kanban systems, etc.).*

AMT favors the Pull system, since the main motivation to invest in this technology is to improve the competitiveness of the organization, such as responsiveness, quality and flexibility [18]. One of the benefits that AMT brings with it when implementing certain technologies is the flexibility to respond to changes in schedules and the combination of products [9], in that sense, the following hypothesis is posed:

> *H_2: AMT has a direct and positive effect on Pull system.*

In addition, with Pull systems and managing the downstream flow, being closer to the market, it has the final authority over how many units to produce [19]. The general philosophy of pull systems is to produce as much as necessary and adapt to external changes faster than push systems and have less inventory accumulation [20]. In that sense, the following hypothesis can be raised:

> *H_3: Pull systema has a direct and poditive effect on Responsive SC strategy.*

2.4 Responsiveness to Market

Responsiveness corresponds to "the ability to respond and adapt effectively over time based on the ability to" read "and understand the real signals of the market" [21].

It corresponds to the speed with which the tasks are performed in which key metrics are used such as order fulfillment, cycle time, delivery cycle time, among others. [22]. Responsiveness in manufacturing is considered as the main source for building responsiveness in the SC [23] and is valued by the following items:

> *Time to market; Delievery speed; Delievery dependability and, manufacturing lead time.*

Currently, AMTs play a fundamental role for the growth of industry and organization, since mass production can be achieved in customer demand in a short time. [24]. In addition, AMT are able to adapt to changes in the variety of products with a short delivery time, while maintaining efficiency and profitability [24]. Since AMT favors production, helps shorten production times and gives companies flexibility, in that sense, the following hypothesis is proposed:

> *H4: AMT has a direct and positive effect on Responsiveness to market.*

Pull-based manufacturing strives to synchronize production with real-time consumption, which increases on-time delivery performance, reduces shortages and reduces costly last-minute changes in orders [25]. A pull system reduces WIP, releases cash flow and space requirements that allow for expansion, quality and less cycle time [26]. Therefore, the following hypothesis can be raised:

> *H5: Pull system has a direct and positive effect on Responsive to market.*

The SC plays a crucial role during the execution of the efficient launch and subsequent product performance [27]. The impact of the SC on the development of new products and the introduction of products is important in areas such as; Fast product delivery to the market, ensuring sufficient inventory in the launch data and ensuring a flow of parts and components for the manufacture of new products [2]. The SC can improve the process of developing new products, reduces development costs and engineering changes, improves product quality and time to market [28]. In that sense, the following hypothesis can be established:H_6: Responsive supply chain has a direct and positive effect on Responsiveness to market. In Fig. 1 the proposed model, as well as the hypotheses, are presented graphically.

3 Methodology

To carry out this investigation, the following steps were carried out:

Step 1. Survey development. A literature review was made in databases such as sciencedirect, emeraldinsight, linkspringer, among others. The search was performed using the keywords; supply chain, supply chain strategy, advance manufacturing techcnology, supply chain responsiveness. The information collected was classified by latent variables according to their affinity. The preliminary questionnaire was submitted to an evaluation by judges. To answer each of the questions, a five-point Likert scale

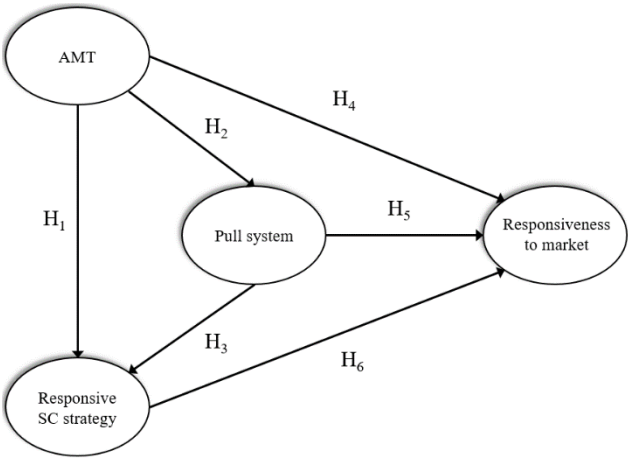


Fig. 1. Proposed model.

Table 1. Industrial sector vs position.

	Industrial sector							
Position	A	B	C	D	E	F	G	Total
Manager	3	0	0	0	0	0	2	5
Engineer	39	2	11	6	4	0	15	77
Supervisor	17	0	8	2	1	0	6	34
Technician	57	2	7	5	1	3	3	78
Operator	27	1	1	1	0	0	1	31
Total	143	5	27	14	6	3	27	225

A Automotive; B Aeronautics; C electric; D Electronics E logistics; F Machining; G Medical

was added where 1 means that the activity is never done and the 5 that the activity is always done.

Step 2. Administration of the questionnaire. The questionnaire is administered in the different sectors of the Mexican maquiladora industry; this was done through a stratified sample to identify the people involved in the activities of interest.

Step 3. Data screening. With the information collected in a database in the SPSS 21® software, a debug is performed to eliminate lost, extreme and unencumbered respondents.

Step 4. Validation of the questionnaire. latent variables are validated through the use of indices such as R^2, Adjusted R^2, composite reliability, Cronbach's alpha, average variance extracted, the inflation index of the variance and Q^2, these indices are proposed by Kock [29].

Step 5. Structural equation model (SEM). SEM is used to validate the hypotheses proposed in Figure 1 and the partial least squares technique integrated in WarpPLS 6.0® software is used, extensively to test relationships with ordinal data, without normality and small samples. The efficiency of the model is evaluated using the Average path coefficient (APC), Average R-squared (ARS), Average adjusted R-squared (AARS), Average block VIF (AVIF), Average full collinearity VIF (AFVIF) and Tenenhaus GoF indices (GoF).

Three effects between latent variables are estimated: direct effects that occur between an independent latent variable and a dependent latent variable and that in Fig. 1 are represented by arrows, the indirect effects that occur between an independent latent variable and a dependent variable a Through a mediating variable and (3) total effects, these are the sum of the previous two. The calculations are made with 95% confidence level and the null hypothesis H_0: $\beta = 0$ is tested vs the alternative hypothesis H_1: $\beta \neq 0$.

4 Results

4.1 Sample Description

254 valid questionnaires were analyzed. Table 1 illustrates that the sector that participated the most was the automotive industry with 143 and the most respondents had the position of technician with 78 participants, followed by engineers with 77. Note that of the 254 questionnaires analyzed, only 225 people gave demographic information.

4.2 Questionaire Validation

Table 2 shows the values of the indices to validate the latent variables and it can be concluded that the questionnaire has enough predictive validity from a parametric point of view since it has R-squared and adj R-squared values are high. Also, the latent variables have content validity, and the Cronbach composite reliability and alpha index are greater than 0.7. Also, the AVE is greater than 0.5 and it is concluded that there is sufficient convergent validity; in addition, there are no multicollinearity problems since the VIF values are less than 3.3 and finally, Q-squared is similar to R^2 and it is concluded that there is non-parametric predictive validity.

4.3 Structural Equation Modeling

Table 3 shows the quality and efficiency indices of the model presented in Fig. 1, it is observed that all the indices have acceptable values. According to the p values associated with the APC, ARS and AARS indices, it is concluded that the model has sufficient predictive validity, there are no collinearity problems and it fits the data, so it is analyzed.

Table 2. Latent variables coefficients.

	AMT	RSCS	Pull System	RtM
R-squared		0.560	0.363	0.718
Adj. R-squared		0.556	0.361	0.715
Comp. Rel	0.908	0.933	0.953	0.942
Conbrach alpha	0.846	0.893	0.901	0.918
AVE	0.768	0.823	0.910	0.802
Full collin VIF	1.730	2.353	3.271	3.224
Q-squared		0.547	0.361	0.705

Table 3. Model fit and quality indexes.

Index	Value	P-value	Index	Value
APC	0.387	< 0.001	AVIF	2.032
ARS	0.547	< 0.001	AFVIF	2.645
AARS	0.544	< 0.001	GoF	0.672

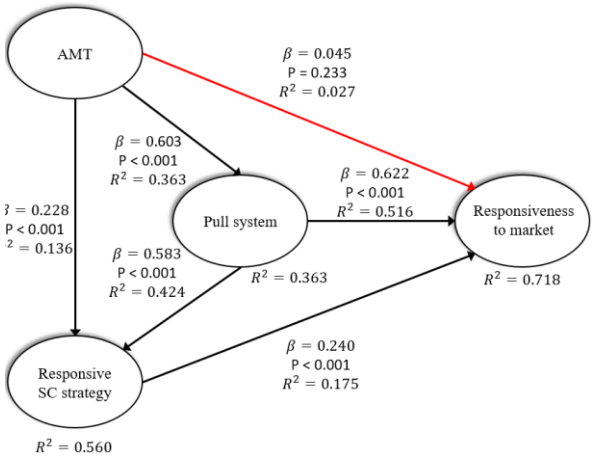


Fig. 2. Evaluated model.

The model evaluated is illustrated in Fig. 2. Based on the p-value associated with β, it is concluded that five hypotheses are statistically significant. For example, for H_1 it is concluded that AMTs have a direct and positive effect on Responsive SC strategy, since when the first variable increases its standard deviation by one unit, the second does so by 0.228 units.

In Fig. 2 a value of R2 is also observed in the dependent variables. For example, 0.718 of Responsive to market is explained in 0.175 by Responsive Supply Chain Strategy (RSCS), in 0.516 of Pull system and in 0.027 by AMT. With these results, it

Table 4. Sum of indirect effects

*ILV	+DLV	Beta	P-value	ES
AMT	RSCS	0.351	< 0.001	0.210
AMT	RtM	0.513	< 0.001	0.302
PS	RtM	0.140	< 0.001	0.116

*Independent latent variable; +Dependent latent variable

Table 5. Total effects

ILV	DLV	Beta	P-value	ES
AMT	RSCS	0.579	< 0.001	0.346
AMT	PS	0.603	< 0.001	0.363
AMT	RtM	0.559	< 0.001	0.329
PS	RSCS	0.583	< 0.001	0.424
PS	RtM	0.761	< 0.001	0.632
RSCS	RtM	0.240	< 0.001	0.175

can be concluded that the Pull system is the most important variable since it explains the 0.516 Responsivenes to market (RtM). Table 4 shows that there are three indirect effects among the variables and that all of them are statistically significant. Table 5 also illustrates the total effects and, all are statistically significant, even though a direct effect was not.

5 Conclusions and Industrial Implications

According to the values obtained in the evaluated model, as well as the variance values of each of the dependent latent variables, the following conclusions can be obtained:

— Regarding the values of the direct effects between the variables, it can be concluded that AMT has a direct effect on the Pull System and on the Responsive SC strategy, not so on Responsiveness to market, but indirectly through the Responsive SC strategy and through the Pull system.

— The results show that having a Pull System within a production system and using tools such as the kanban system, reducing batches and setup times help companies cope with sudden changes in the market, also help reduce manufacturing times and therefore, delivery times, but above all, that there is reliability in delivery. In that sense, managers must ensure that the Pull System runs correctly.

— The results also show that AMTs indirectly help companies to be able to respond effectively to offer a rapid response in the changing market through the implementation of a pull system.

References

1. Hum, S.H., Parlar, M., Zhou, Y.: Measurement and optimization of responsiveness in supply chain networks with queueing structures. European Journal of Operational Research, 264(1), pp. 106–118 (2018)
2. van Hoek, R.: From tinkering around the edge to enhancing revenue growth: supply chain-new product development. Supply Chain Management: An International Journal, 11(5), pp. 385–389 (2006)
3. Pero, M.: A framework for the alignment of new product development and supply chains. Supply Chain Management: An International Journal, 15(2), pp. 115–128 (2010)
4. Gules, H.K., Burgess, T.F.: Manufacturing technology and the supply chain: Linking buyer-supplier relationships and advanced manufacturing technology. European Journal of Purchasing & Supply Management, 2(1), pp. 31–38 (1996)
5. Narasimhan, R., Kim, S.W., Tan, K.C.: An empirical investigation of supply chain strategy typologies and relationships to performance. International Journal of Production Research, 46(18), pp. 5231–5259 (2008)
6. Koufteros, X.A.: Testing a model of pull production: a paradigm for manufacturing research using structural equation modeling. Journal of Operations Management, 17(4), pp. 467–488 (1999)
7. Díaz-Reza, J.R., García-Alcaraz, J.L., Gil-López, A.J., Blanco, J.: Design, process and commercial benefits gained from AMT. Journal of Manufacturing Technology Management (2019)
8. Krot, K., Mazgajczyk, E., Risińska, M., Woźna, A.: Strategy of Improving Skills of Innovation Managers in the Area of Advanced Manufacturing Technologies. Intelligent Systems in Production Engineering and Maintenance, Cham: Springer International Publishing (2019)
9. Sohal-Amrik, S.: Implementing advanced manufacturing technology: Factors critical to success. Logistics Information Management, 5(1), pp. 39–46 (1992)
10. Aravindan, P., Punniyamoorthy, M.: Justification of Advanced Manufacturing Technologies (AMT). The International Journal of Advanced Manufacturing Technology, 19(2), pp. 151–156 (2002)
11. Kaplan, R.S.: Must CIM be justified by faith alone? (1986)
12. Siha, S., Linn, R.J.: A zero-one goal programming decision model for selecting technology alternatives. In: International Industrial Engineering Conference (1989)
13. Lyons, A.C., et al.: The development of supply chain strategy. In Customer-Driven Supply Chains: From Glass Pipelines to Open Innovation Networks, Lyons, A.C., et al., Editors, Springer London: London, pp. 1–19 (2012)
14. Munksgaard, K.B., Stentoft, J., Paulraj, A.: Value-based supply chain innovation. Operations Management Research, 7(3), pp. 50–62 (2014)
15. Birasnav, M., Bienstock, J.: Supply chain integration, advanced manufacturing technology, and strategic leadership: An empirical study. Computers & Industrial Engineering, 130, pp. 142–157 (2019)
16. McDermott, C.M., Greis, N.P., Fischer, W.A.: The diminishing utility of the product/process matrix: a study of the US power tool industry. International Journal of Operations & Production Management, 17(1), pp. 65–84 (1997)
17. Li, J.W.: Investigating the efficacy of exercising JIT practices to support pull production control in a job shop environment. Journal of Manufacturing Technology Management, 16(7), pp. 765–783 (2005)

18. Burgess, T.F.: Supply-chain collaboration and success in technology implementation. Integrated Manufacturing Systems, 8(5), pp. 323–332 (1997)
19. Baykal-Gürsoy, M., Altiok, T., Danhong, H.: Look-back policies for two-stage, pull-type production/inventory systems. Annals of Operations Research, 48(4), pp. 381–400 (1994)
20. Altiok, T.: Pull-Type Manufacturing Systems. In Performance Analysis of Manufacturing Systems, Altiok, T. Editor. Springer New York, pp. 273–351 (1997)
21. Catalan, M., Kotzab, H.: Assessing the responsiveness in the Danish mobile phone supply chain. International Journal of Physical Distribution & Logistics Management, 33(8), pp. 668–685 (2003)
22. Ross, D.F.: Crafting business and supply chain strategies, in distribution planning and control: managing in the era of supply chain management, Ross, D.F. Editor. Springer New York, pp. 83–140 (2015)
23. Sandberg, E.: Retail supply chain responsiveness. International Journal of Productivity and Performance Management, 67(9), pp. 1977–1993 (2018)
24. Nath, S., Sarkar B.: Performance evaluation of advanced manufacturing technologies: A De novo approach. Computers & Industrial Engineering, 110, pp. 364–378 (2017)
25. Kumar, S.: Achieving customer service excellence using Lean Pull Replenishment. International Journal of Productivity and Performance Management, 62(1), pp. 85–109 (2013)
26. Aghazadeh, S.M.: Does manufacturing need to make JIT delivery work? Management Research News, 27(1/2), pp. 27–42 (2004)
27. Kou, T.-C.: The influence of supply chain architecture on new product launch and performance in the high-tech industry. Journal of Business & Amp, Industrial Marketing, 30(5), pp. 677–687 (2015)
28. Ragatz, G.L., Handfield, R.B., Petersen, K.J.: Benefits associated with supplier integration into new product development under conditions of technology uncertainty. Journal of Business Research, 55(5), pp. 389–400 (2002)
29. Kock, N.: WarpPLS user manual: Version 6.0. ScriptWarp Systems. Laredo, TX, USA (2017)

ISSN 1870-4069

A Sentiment Analysis Approach for Drug Reviews in Spanish

Karina Castro Pérez[1], José Luis Sánchez Cervantes[2], María del Pilar Salas Zárate[1], Luis Ángel Reyes Hernández[1], Lisbeth Rodríguez Mazahua[1]

[1] Tecnológico Nacional de México,
Instituto Tecnológico Orizaba,
Mexico

[2] CONACYT,
Instituto Tecnológico de Orizaba
Mexico

karinacastro.058@gmail.com, jlsanchez@conacyt.mx,
{msalasz, lreyes, lrodriguezm}@ito-depi.edu.mx

Abstract. The analysis of opinions in the medical context is of great relevance for health care since it allows us to gain insight from the experiences and opinions of patients and health professionals regarding nutrition, exercise, and other health related issues. The rise in the application of opinion mining in recent years is a direct consequence of the growth of social networks and blogs that generate a large volume of unstructured data, however, the manual review of such data is not feasible due to the amount that is generated in real time. Thus, opinion summarization systems that use Web Scraping techniques and opinion mining are needed. In this sense, this work presents a solution proposal under a hybrid approach based on both semantic approach and machine learning approach for the development of an opinion mining analysis system applying Web Scraping and Natural Language Processing (NLP) techniques to know the users' experiences about drugs for chronic-degenerative diseases available in blogs, video blogs and specialized websites in Spanish language.

Keywords: Web scraping, natural language processing, opinion mining, Spanish language analysis, drugs opinion.

1 Introduction

Opinion mining is an area of great importance for the coarse application that has, focuses on analyzing opinions, sentiments, evaluations, assessments, attitudes and emotions of people towards entities such as products, services, organizations, individuals, problems and events. [1]; the application of this field has increased over the years as a result of the growth of social networks and blogs that generate vast volume amounts of unstructured data, nevertheless, the manual revision of the data is not feasible due to the amount that is generated in real time.

Therefore, it is necessary to apply Web Scraping and opinion mining techniques that allow summarizing the information and obtaining precise knowledge, of interest in a certain area, which will result useful in a decision making process.

According to the World Health Organization (WHO) in 2016 [2], diabetes mellitus, hypertension, cardiovascular diseases, cancer, among others diseases catalogued as chronic-degenerative, are positioned among the top ten causes of mortality in Mexico and worldwide; as a result, health systems need to analyze important aspects such as eating habits, exercise and treatments. In this context, the application of opinion mining is valuable because it allows analyzing the comments of patients and health professionals to identify symptoms and drugs related to chronic-degenerative diseases. In addition, it allows health specialists to speed up the process of identification and selection of the drugs that they prescribe, allowing them to dedicate more time to the physical exploration of the patient to prevent additional complications to the disease, which results in higher quality attention for the patient. The main contribution of this work is a hybrid approach that applies semantics through a tagged corpus and supervised machine learning for an opinion analysis system for drugs, searching blogs, video blogs and specialized websites, in the Spanish language, implementing Web Scraping techniques and Natural Language Processing (NLP).

This paper is structured as follows: Section 2 presents a set of recent works related to our proposal. It describes initiatives of healthcare-oriented opinion mining and a comparison of articles that address polarity detection as an opinion mining activity, the use of NLP and the sources from which the data were obtained. The proposed hybrid approach for the opinion analysis system for drugs in Spanish based on Web Scraping techniques is presented in Section 3; Section 4 presents a case study for polarity analysis in comments published in Spanish language on drugs for chronic-degenerative diseases. Finally, the Section 5 presents the conclusions and future work.

2 Related Work

An analysis that describes the difficulty of finding Adverse Drug Events (ADE) in clinical trials conducted by pharmaceutical organizations was developed by Lee et al. [3]. For this reason, authors examined deep learning models, however, it was found that employing such a method is costly given the scarcity of Twitter ADE tweets. Similarly, a method to extract ADE of Twitter, through a five-step channel was proposed in [4]. Such steps are: 1) Tweet capture; 2) Data pre-processing; 3) Drug-related classification; 4) Tweet sentiment analysis and, 5) ADE extraction from Twitter data. The potential of sentiment analysis in medicine with data from clinical narratives and medical social networks was described by Denecke and Deng [5].

Through a literature review, we summarize linguistic peculiarities of sentiment in medical texts. On the other hand, in [6], an investigation of the NLP techniques for the extraction of opinions and sentiment analysis was carried out, which identified the stages of pre-processing required to structure the texts. It was found that tokenization is an essential task for Chinese and Japanese, among others, because their words are composed differently from the Latin alphabet.

Table 1. Comparative of related works.

Initiative	Approach	Polarity detection	Data Source	NLP	Language
Lee et al. [3]	Semi-supervised convolutional neural network.	✓	Twitter	X	English
Y. Peng et al. [4]	Supervised machine learning and linguistic method.	✓	Twitter	✓	English
Denecke & Deng [5]	-----	✓	Clinical narratives, social networks and specialized websites.	X	English
Solangi et al. [6]	-----	✓	-----	✓	English, Chinese, Japanese, Arabic, French, Spanish, and German
Salas-Zárate et al. [7]	Semantic	✓	Twitter	✓	English

A method for sentiment analysis that detects diabetes-related aspects in tweets, using ontologies to semantically describe the relationships between concepts in the specific domain. Salas-Zárate et al. [7] presented a proposed sentiment classification approach divided into three main components: the pre-processing module for cleaning and correcting text, the semantic annotation module for aspect detection, and the sentiment classification module that calculates polarity.

As shown in Table 1, the analyzed works use one of several approaches, such as the application of algorithms for the classification of sentiment and the use of semantic methods, both of which are used in our system as part of a hybrid approach. Unlike [3] and [8], our approach gets comments from a variety of sources such as forums, blogs and video blogs, where there is relevant information that has not been analyzed in detail, so the scarcity of comments does not prove to be an obstacle as it is with getting tweets about medication from the social network Twitter.

Our approach makes a review on chronic-degenerative diseases, so a vast amount of comments is obtained through Web Scraping, and they contain information about sentiments regarding a drug, the dosage of the drug, the price and even the adverse effects that it has on patients.

On the other hand, a hybrid approach is proposed in [4] using sentiment classification algorithms such as Syntactic dependency paths, in addition, linguistic method through external tools, unlike our approach, we adopt the supervised machine learning that integrates a semi-automatic tagged corpus that makes use of a dictionary of positive and negative words to tag the corpus, to ensure that the corpus is correct,

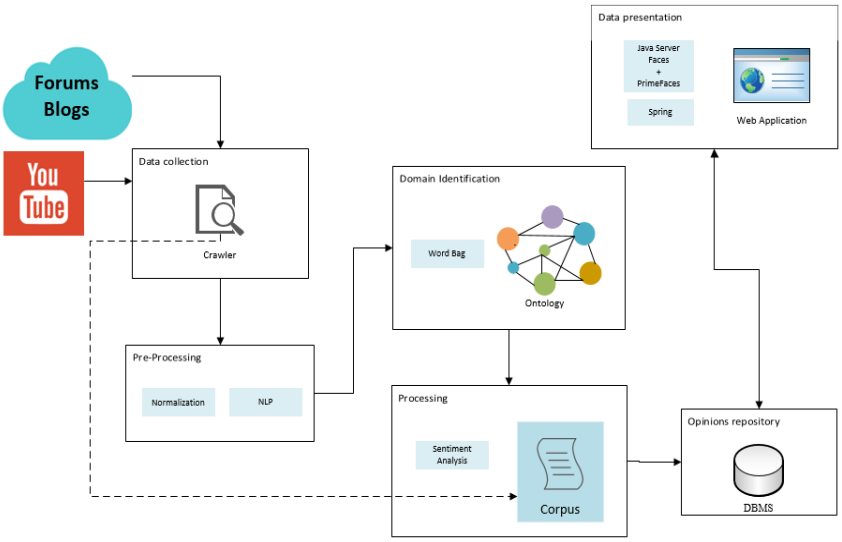


Fig. 1. Architecture of *SentiScrap* system.

two health specialists analyzed the comments collected and tagged, where they found mentions of diseases, medications and symptoms and the use of automatic learning. In addition, our approach gets first-hand comments from patients unlike the clinical narratives discussed in [5] which are subject to interpretation by third parties, by doctors and nurses, so they are less accurate. Finally, as mentioned in [7], very scarce work has been done in the implementation of opinion mining and NLP in languages other than English, such as Chinese and Japanese, among others, because it requires a great effort for implementation and analysis, while our approach addresses opinion mining and NLP for comments in the Spanish language as it is recognized as being the second most spoken in the world.

3 Approach

In order to achieve a successful outcome, a hybrid approach using semantics and automatic learning is proposed for a Spanish opinion analysis system for chronic-degenerative disease drugs, based on Web Scraping techniques, consisting of six main modules and a corpus: 1) Data collection module; 2) Pre-processing module; 3) Domain identification module; 4) Processing module; 5 Opinions repository, and 6) Data presentation. Figure 1 shows the architecture of system. In this approach, a corpus was created with 280 comments obtained from forums on diseases such as type 2 diabetes mellitus, hepatitis and hypertension. Given that polarity detection is limited in the Spanish language due to the lack of data sets, we made a corpus tagged semi-automatically with the use of a dictionary that was manually built, from the analysis of comments found in specialized forums, blogs and video blogs to identify the words that tell a positive comment from a negative one. To ensure the effectiveness of corpus

labeling, two health specialists were consulted; they manually reviewed each comment to corroborate that both, the positive and negative comments presented the corresponding tag.

3.1 Data Collection Module

In this module, we collected reviews on drugs and symptoms of chronic-degenerative diseases from forums, blogs, and video blogs. Specifically, diabetes, hypertension, and hepatitis diseases were considered.

3.2 Pre-Processing Module

The pre-processing of data is an important step for the normalization of the text, therefore, we chose to use three phases for the treatment of our data:

1. Delete unusual characters: comments contain special characters that do not provide information, so they are deleted.
2. Delete duplicate comments: This step is important, because duplicate comments affect the final result of the analysis, therefore it is important to ensure that duplicate comments are removed.
3. Delete comments that only have URLs: comments that only include links to other sites do not contribute as a comment to analyze polarity, for this reason they are discarded.

The application of these tasks on the comments ensures a better analysis of sentiments, however, the incorrect use of language is a common scenario, created by the use of abbreviations or spelling errors on behalf of the users, requiring a greater effort to carry out opinion mining activities, for that reason, this module also makes use of a spell checker.

3.3 Domain Identification Module

The identification of words has high relevance because it allows the verification that the comments obtained is the mention of a drug prescribed for diabetes, hypertension and hepatitis, this provides as a result a set of data with valuable information for analysis. Therefore, a bag of words is used, made through an ontology of medical domain based on Snomed [8].

3.4 Processing Module

This module adopts the supervised automatic learning approach which makes use of a semi-automatic labeled corpus, necessary to train the algorithm that performs the sentiment analysis, this permits it to recognize new opinions in the Spanish language and to classify them correctly.

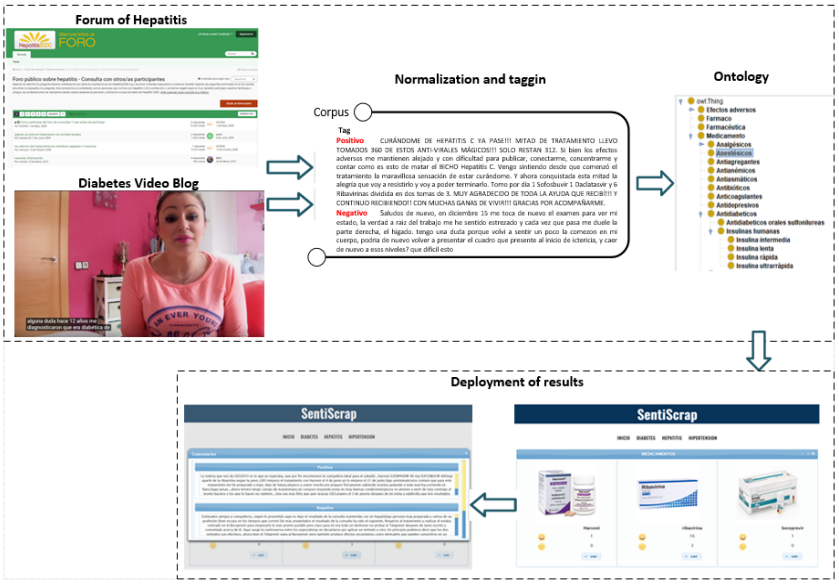


Fig. 2. *SentiScrap* Workflow.

3.5 Opinions Repository Module

The opinions and polarity of the drugs resulting from the analysis are stored in a database to keep the data available for consultation.

3.6 Data Presentation Module

A web interface is presented that supports the interaction of health specialists with the "*SentiScrap*" system to know the comments and polarity of the drugs prescribed for chronic-degenerative diseases, specifically diabetes mellitus, hepatitis and hypertension.

4 Case Study: Polarity Analysis in Comments Published in Spanish Language on Drugs for Chronic-Degenerative Diseases

For this case study, the medical domain of application was selected, specific for chronic-degenerative diseases. Suppose that a health specialist needs to know quickly the opinion of patients undergoing treatment to alleviate hepatitis C disease:

— How will the specialist get patient feedback in Web forums in Spanish language?
— How will the health specialist know which comments have a positive or negative impact?

Fig 3. *Hepatitis* options of the *SentiScrap* system

— How will the specialist be able to identify the best medications prescribed for the treatment of hepatitis C?

As a solution, it is intended that the health specialist has access to the information analyzed through the "*SentiScrap*" system. Figure 2 describes the workflow for user interaction with the system's functionalities.

Suppose that doctor wants to know the prescription drugs for hepatitis C, those that patients comment on in the forums. The user selects the word "Hepatitis" in the menu options, the system generates a query to show the medicines associated to the disease, and the number of positive and negative comments of each of these drugs. Figure 3 shows the results of the polarity analysis on three prescription drugs for hepatitis, Harvoni, Ribavirina and Boceprevir. The polarity reveals that Ribavirina has 16 positive and 3 negative opinions, while Harvoni and Boceprevir have the same number of positive and negative opinions. These results, provide information to the doctor to know the drug that receives better opinions from users.

In addition, the user would like to know the comments on each drug, which, by clicking on the "ver" button, this action opens a modal window with each comment

Fig 4. *SentiScrap* comments modal window

referring to the drug and its classification with respect to whether it is a positive comment or a negative one. Figure 4 shows a modal window with the collected comments.

The proposed solution aims to show that the information contained in the forums and blogs is of great relevance to health specialists, because by accessing the comments of patients, the doctor can know the experience of each patient, thus consulting first-hand data, which are useful for decision making.

5 Conclusions and Future Work

The analysis of the related work shows the existence of great opportunity in the application of studies related to the analysis of opinions and polarity detection in specialized websites, blogs or video blogs. For this reason, we propose a hybrid approach, through supervised machine learning and the use of semantics through a tagged corpus. The approach allows the analysis of the sentiments of the opinions for the drugs prescribed for chronic-degenerative diseases in a successful way, reducing time and effort in the search for relevant information on diabetes, hypertension and hepatitis diseases.

As future work, it is contemplated to incorporate more aspects to the opinions analysis, such as the adverse effects to the medicines, time of the treatment and price, likewise, we add the information to the web application, through graphs for a better visualization and analysis of the information, allowing to improve the taking of decisions that the specialists of the health make. In addition, we consider adding a module to validate users' opinions about medicines. This validation consists to rely of

experts in the medical domain, such as cardiologists, nephrologists, among others, verifying that the opinions made by users are adequate, mainly to avoid self-medication.

Acknowledgments. This research work was sponsored by the National Council for Science and Technology (CONACYT) and the Secretariat of Public Education (SEP). The authors are grateful to Tecnológico Nacional de México (TNM) for supporting this work.

References

1. Liu, B.: Sentiment Analysis and Opinion Mining, pp. 168 (2012)
2. WHO: Las 10 principales causas de defunción. https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death (2019)
3. Lee, K., Qadir, A., Hasan, S.A., Datla, V., Prakash, A., Liu, J., Farri, O.: Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In: Proceedings of the 26th International Conference on World Wide Web, pp. 705–714 (2017)
4. Peng, Y., Moh, M., Moh, T.S.: Efficient adverse drug event extraction using twitter sentiment analysis. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1011–1018 (2016)
5. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine, 64(1), pp. 17–27 (2015)
6. Solangi, Y.A., Solangi, Z.A., Aarain, S., Abro, A.G., Mallah, A., Shah, A.: Review on Natural Language Processing (NLP) and its toolkits for opinion mining and sentiment analysis. In: IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1–4 (2018)
7. Salas-Zárate, M. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M.Á., Valencia-García, R.: Sentiment analysis on tweets about diabetes: An aspect-level approach. Computational Mathematical Methods in Medicine, 2017(5), pp. 1–9, (2017)
8. NCBO BioPortal. https://bioportal.bioontology.org/ontologies (2019)

ISSN 1870-4069

An Architecture for an IoT-based Telecare System for the Elderly Using Big Data Analytics

Jesús Miguel Echevarría Díaz[1], José Luis Sánchez Cervantes[2], Luis Omar Colombo Mendoza[1], Giner Alor Hernández[1], Ignacio López Martínez[1]

[1] Tecnológico Nacional de México,
Instituto Tecnológico Orizaba,
Mexico

[2] CONACYT,
Instituto Tecnológico de Orizaba,
Mexico

{jmechevarria2015, nachocero}@gmail.com, jlsanchez@conacyt.mx,
{lcolombom, galor}@ito-depi.edu.mx

Abstract. The accelerated aging of the population is a real and constant situation worldwide. Long-lived population is the most affected demographic sector in terms of health and home safety. The advent of Big Data, in conjunction with the Internet of Things (IoT) omnipresence, have made telecare-based systems an increasingly desirable alternative when dealing with the treatment and remote care of the elderly and frail, allowing them to go on with their lives in a normal, almost independent manner. This situation has spawned an ever increasing demand for innovative telecare solutions. In order to present an alternative solution to the problem at hand, this paper proposes a telecare system architecture based on IoT and Big Data Analytics (BDA) to monitor the user's vital signs and activities, send alerts to the caregivers whenever necessary, with the goal of detecting and preventing accidents in the home.

Keywords: big data analytics, internet of things, telecare, wearable, decision tree.

1 Introduction

The impact of BDA techniques in healthcare organizations shows promising research directions, especially when implementing novel use cases for potential healthcare frameworks. Over 10 million people in Mexico are 60 years of age or more, frequently spending long periods of time alone at home; which increases the probabilities of going through a medical emergency while on their own.

Table 1. Comparative analysis of related papers.

Authors	Usage of sensors or wearables	Prevention	Big Data Analytics	Contribution
Yacchirema et al. [1]	Both	×	✓	System for the detection and support of treatment of Obtrusive Sleep Apnea of elderly people
Yacchirema et al. [2]	Both	×	✓	A decision tree-based Big Data model for fall detection
Jankowski, Schönijahn and Wahl [3]	None (computer vision)	✓	×	Research on tele-monitoring
Khalifeh et al. [4]	Sensors	×	×	Build an e-health monitoring and fall detection system
Meissner [5]	Both	×	×	Build a fall detection system that monitors in real-time an older adult
Silapachote et al. [6]	Both	×	×	Fall detection and response tracking application
Walters et al. [7]	Sensors	×	×	Design and test modules of a monitoring system based on multiple sensor feedback in a bathroom setting

Two main approaches for this type of problem were identified during the analysis of the literature, namely, computer vision techniques using image-recognition and streaming, and a sensor-based approach using wearable devices to monitor the user's physical health.

All telecare systems proposed in the literature focus on the detection of accidents, while ours focuses on preventing them using predictive analysis; a rule-based classification algorithm was implemented using a decision tree as part of the architecture proposed in this paper, hinging on the use of IoT devices and BDA techniques to prevent possible accidents in the home.

After carrying out the case study, it was found that the use of big data analytics on data coming from wearable devices provides a viable solution to the problem stated, making it possible to send notifications to any interested party, alerting about changes in the user´s physiological parameters which could result in a dangerous, and otherwise unforeseen, situation, allowing caregivers to act accordingly and in a prompt manner.

The rest of the paper is distributed as follows. Section 2 contains a comparative analysis of related work in the field of IoT and BDA in telecare systems. In section 3, the system's architecture and its modules are described. Section 4 presents a case study, and finally, conclusions and future work are included in section 5.

2 Related Work

Table 1 shows a concise description of the most relevant features of the scientific papers included as related work.

As shown in Table 1, using wearable sensors as a data source for innovative telecare systems for fall detection and indoor monitoring of frail or elderly people, is an approach which constantly gains preference in IoT and/or Big Data contexts. The telecare systems proposed by the authors focus primarily in the detection of a falls, using sensors [4, 7], as well as wearables [1, 2, 5, 6], while [3] presents a research on tele-monitoring for the prevention of accidents. In this aspect our work differs from the rest in that it proposes a telecare system architecture to prevent accidents in the home by monitoring physiological parameters obtained via a wearable device and using them in conjunction with the user's age and gender to send alerts when a set of irregular values are recorded during a relatively short period of time.

Our architecture is comprised of several modules. Data will be retrieved through constant monitoring of a wearable device manufactured by Fitbit Inc., through its API endpoints. All collected data will be stored in the non-relational database management system *Apache Cassandra™* and fed to an *Apache Spark™* cluster for analytical purposes, with the goal of preventing accidents in the home, generating alerts and notifications sent to the user's caregivers and family members via mobile and web applications.

3 Architecture

Fig. 1 shows the proposed multi-tiered architecture and its modules. The function of each module is briefly described below.

3.1 IoT Layer

Data acquisition module. Data is periodically collected through Fitbit's API endpoints in JavaScript Object Notation format, including, but not limited to, the user's vital signs, burned calories and physical activities, allowing the system to assess, to a certain extent, the user's physical health and whether or not there is a potentially dangerous situation, alerting the caregivers and family members accordingly.

3.2 Storage Layer

Big Data repository. Data coming from the IoT layer will be stored in a Cassandra database cluster, which will serve as a data frame provider for the classification algorithms run in the Spark cluster.

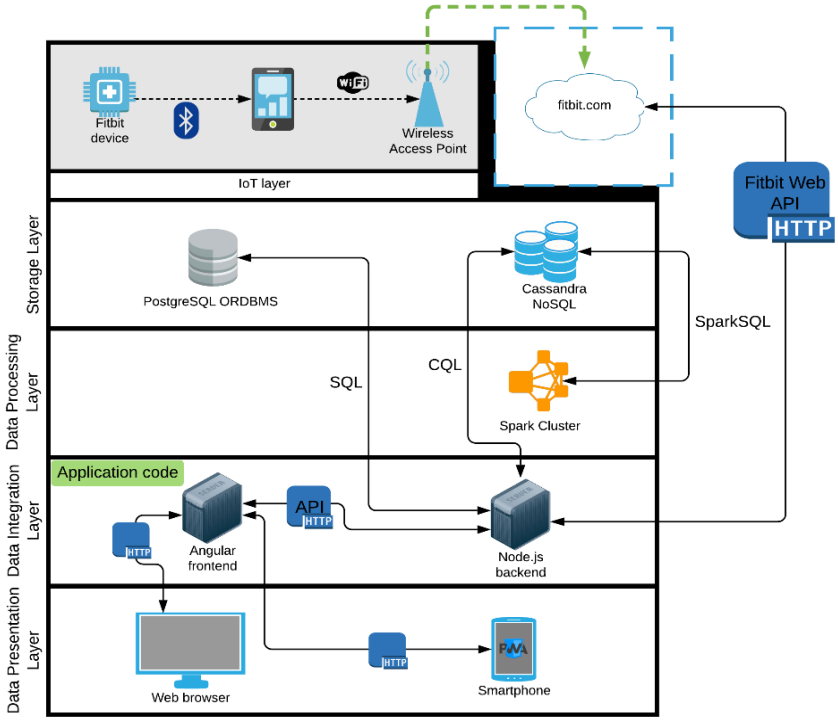


Fig. 1. Architecture.

PostgreSQL repository. The PostgreSQL Object Relational Database Management System will be used to store any information, which does not come from the IoT layer, thus not being used in any Big Data Analytics process, but still needed for the complete functioning of the system.

3.3 Data Processing Layer

This layer contains the Spark cluster, in which machine learning algorithms will be carried out, using a predictive model designed to detect patterns and anomalies in the user's physical health and activities, with the purpose of preventing accidents. Parameters like sleep status, heart rate levels and physical activity, the latter in the form of walked distance or steps taken, will produce the input variables for the algorithms, and the result will be a notification or alert of the elder's condition.

3.4 Data Integration Layer

Backend module. All data collected will be transmitted via web services and into the applications' backend module in the Node.js server. The server will persist the data in

Date	Out of Range	Fat Burn	Cardio	Peak
2019-09-14	721.55 cal. 529 min.	251.06 cal. 77 min.	0.00 cal. 0 min.	15.19 cal. 1 min.
2019-09-13	392.57 cal. 277 min.	2,929.92 cal. 743 min.	97.68 cal. 10 min.	42.97 cal. 3 min.
2019-09-12	1,070.04 cal. 741 min.	1,332.58 cal. 492 min.	33.71 cal. 4 min.	0.00 cal. 0 min.
2019-09-11	1,023.86 cal. 755 min.	1,694.78 cal. 525 min.	579.42 cal. 58 min.	36.80 cal. 3 min.
2019-09-10	1,289.98 cal. 877 min.	1,459.28 cal. 499 min.	415.67 cal. 47 min.	75.58 cal. 6 min.
2019-09-09	1,190.57 cal. 755 min.	2,163.67 cal. 626 min.	48.66 cal. 5 min.	0.00 cal. 0 min.

Time	Heart rate
01:31:00	95
01:32:00	93
01:33:00	92
01:34:00	92
01:35:00	94
01:36:00	94

Fig. 2. Heart rate information as it is presented in the web application.

Cassandra for later processing, and, in occasions, will also send it directly to the frontend module to be presented to the applications' users. This distinction is made according to what triggered the data transmission process, an alert from the system or a normal request from the user.

Frontend module. The frontend module will only make requests directly to the backend, retrieving all required information to be shown to the users in the presentation layer via the web and mobile applications.

3.5 Data Presentation Layer

Information like reports and records will be constantly accessible to caregivers and family members through the web and mobile applications. Alerts and notifications will be sent according to the results obtained from the algorithms executed in the backend and data processing modules. Such events will be triggered as a consequence of a dangerous situation being detected by the system, or an accident which already occurred and requires immediate action on behalf of the applications' users, namely the relatives and wardens.

4 Case Study

A rule-based classification algorithm was implemented using a decision tree. The algorithm uses the data coming from the Fitbit device, namely the user's age, sex and heart rate. Fig. 2 shows two tables with heart rate data, one contains daily aggregates (left), and the other, a per-minute representation (right) of the date corresponding to the highlighted row, as it is presented to the user in the web application.

Fig. 3 illustrates a graphical representation of the algorithm, a sample input for a given instance of time, and the resulting information shown to the user; input

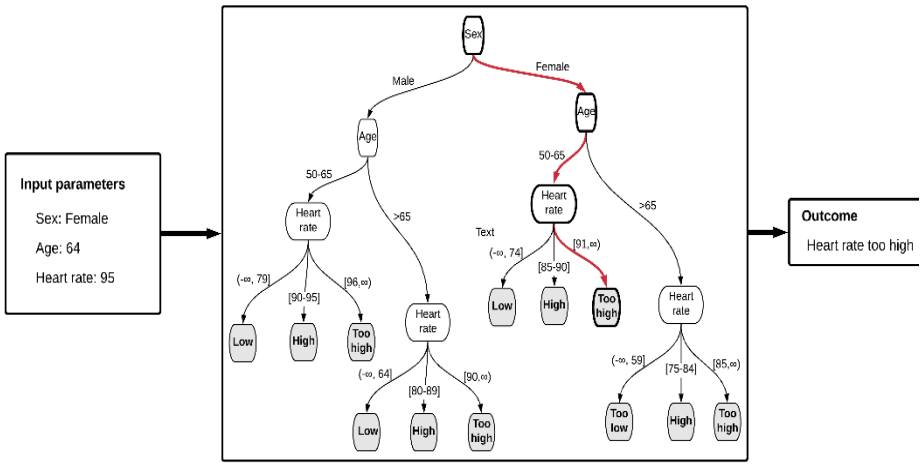


Fig. 3. A graphical representation of the algorithm.

parameters correspond to a 64-year-old female subject, whose heart rate, at a given instance of time, is 95, which is considered too high for this input set.

One abnormal value is not a direct indicator of a problem, for example, the simple action of standing up can cause a temporary surge in heart rate, even more when dealing with older persons, which is why the alert is not sent on every abnormal value detected. Heart rate data is requested every ten minutes to the Fitbit Web API, containing values from the last time data was received until present time, the algorithm then executes using the dataset corresponding to the last hour; results are then grouped for every five consecutive minutes, and if in any of those groups, abnormal values represent at least 80%, a notification is sent to the caregivers, so they are aware of the situation and are able to take appropriate actions.

5 Conclusions and Future Work

The continuous aging of the population, alongside the increase of chronic diseases, can cause home accidents to occur more frequently, especially when involving the elderly; this situation can be tackled effectively through the development and installation of telecare systems, taking advantage of the usage of wearable devices, sensors and surveillance devices, and by harnessing the computing capabilities of BDA tools. Taking this into consideration, a BDA and IoT based telecare system was proposed in this paper, to monitor the user's vital signs, and send alerts on potentially dangerous situations, with the goal of preventing accidents in the home. Our case study verified the effectiveness of the decision tree approach as part of the architecture to prevent accidents in an indoor environment, quickly making caregivers and family members aware of the situation.

Future work will include the expansion of the model used in the decision tree as part of a new iteration of the KDD (Knowledge Discovery in Databases) process,

because the current one, although effective in the detection of abnormal heart rate values on a per-second basis, can be improved by using additional parameters, such as age, weight, sleep status and steps taken in a period of time; the goal is to factor the user's physical and mental exhaustion into the algorithm, which would yield more accurate predictions regarding a possible fall.

Also, a data source will be added in the IoT layer, in the form of a Computer Vision module, comprised of a camera device, attached to a Raspberry Pi B+ board. A computer vision software will be programmed using OpenCV, and will be installed in the board. This module will add up to the wearable's input in the data acquisition process, this time including the objects layout as well as the elder's position in the room. A prediction model will be designed and implemented in Spark, which, with the added data source, would effectively increase the accuracy of the predictions.

Acknowledgements. The authors are very grateful to the National Institute of Technology of Mexico (TecNM) for supporting this work. We would also like to thank the sponsors of this paper: the National Council of Science and Technology (CONACYT), and the Secretariat of Public Education (SEP) through the Program for Professional and Educational Development in Higher Studies (PRODEP).

References

1. Yacchirema, D.C., Sarabia-Jácome, D., Palau, C.E., Esteve, M.: A smart system for sleep monitoring by integrating IoT with big data analytics. IEEE, 6, pp. 35988–36001 (2018)
2. Yacchirema, D., de Puga, J.S., Palau, C., Esteve, M.: Fall detection system for elderly people using IoT and Big Data. Procedia Computer Science, 130, pp. 603–610 (2018)
3. Jankowski, N., Schönijahn, L., Wahl, M.: Telemonitoring in home care: creating the potential for a safer life at home. In: Safe at Home with Assistive Technology, Springer, pp. 81–93 (2017)
4. Khalifeh, A., Saleh, A., Al-Nuimat, M., Tair, D.A.: An open source cloud based platform for elderly health monitoring and fall detection. In: Proceedings of the 4th Workshop on ICTs for Improving Patients Rehabilitation Research Techniques, pp. 97–100 (2016)
5. Meissner, E.: Wearable Based Fall Detection System for the Elderly (2017)
6. Silapachote, P., Srisuphab, A., Phongpawarit, J., Visetpalitpol, S., Jirapasitchai, S.: REDLE: a platform in the cloud for elderly fall detection and push response tracking. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 10(2), pp. 185–195 (2016)
7. Walters, T., Espinal, R., Restrepo, V., Santacroce, J., Dow, D.: Activity monitoring system for independent Elderly living. In: Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), pp. 116–119 (2016)

ISSN 1870-4069

A Process for Automatic Generation of Medical Mobile Applications using Voice Recognition

Jesús Fernández Avelino, Giner Alor Hernández, Mario A. Paredes Valverde, Lisbeth Rodríguez Mazahua, María A. Abud Figueroa

Tecnológico Nacional de México,
Instituto Tecnológico de Orizaba,
Mexico

{fdezjesus12, mapv1015, mabudfig}@gmail.com,
{galor,lrodriguezm}@ito-depi.edu.mx

Abstract. In software engineering, the adoption of mobile devices generates the need for proposing new methods and processes for developing mobile applications. This paper presents and describes a software development process for the automatic generation of medical mobile applications using voice recognition. The software development process is supported by a conversational agent that captures the functional requirements of the application to be developed. A Web-based prototype was developed as a proof-of-concept of the process proposed. We believe this process will allow reducing time and effort to developers in mobile applications.

Keywords: voice recognition, design patterns, software development process.

1 Introduction

Over the years mobile phones have changed from a simple communication device to an operative tool. People perform tasks from a mobile application, this is made possible through the development of mobile applications. Now these mobile applications have become an integral part of our lives and we rely on them in more than one way. Nowadays there are a lot of mobile healthcare applications that allow registering medical data in order to prevent, control, and monitoring diseases such as diabetes and hypertension. These applications use voice recognition which allows optimizing time and effort.

Voice recognition is the ability of a computer to understand and execute voice commands. Voice recognition is a way of stablishing communication between user and computer, it aims to replace other ways of interaction such as a mouse, keyboard, touch functionality to mention but a few. Nowadays, it is possible to use mobile devices, cars, smart TVs and domestic automation technology by voice recognition in order to perform everyday tasks. Nevertheless, this kind of interaction is used in different contexts. Voice recognition is very important for the business and professional sectors.

This paper presents and describes a process to automatically generate user interfaces of mobile healthcare applications by using voice recognition. This paper is organized into four sections, Section 2 presents related work, Section 3 describes the process to generate UI automatically, Section 4 a prototype as a proof-of-concept of the process proposed. Finally, conclusions and remarks are provided.

2 Related Work

The following is a summary of the most important related work according to voice recognition, which allowed it obtains valuable information related to this work.

Erić et al. [1] carried out a detailed comparison between different voice recognition tools such as Jasper platform, Google speech API (Application Programming Interface), Alexa Voice Service and Bing speech API to implement them in home automation projects. Moreover, Cortes et al. [2] presented a generator of mobile applications based on User Interface Design Patterns named Atila. In contrast to others, Atila generates a native project depending on the specified platform and it allows modifying the project according to user needs. In another research, Joon et al. [3] analysed the voice recognition vulnerabilities of mobile devices. Also, an attack model named Toilet-time was presented and tests were performed by using BadVoice tool in order to simulate a hands-free cell phone and execute voice commands. Modak et al. [4] developed a desktop application that implements an interface of natural language processing to generate Web pages, the implementation results are promising, nevertheless, voice recognition and speech-to-text process depend on factors such as environment, environmental noise, microphone quality to mention but a few.

Moreover, in the user interface design approach Vittone et al. [5] suggested using *wireframes* which are a simplified representation of an individual screen and help to have a better idea of the screen's layout and identify informative and interactive elements. Beltramelli [6] presented a software based on convolutional neuronal networks, which generates source code for Web and mobile platforms from an image and a previous training model, the software is named pix2code. On the other hand, Patil et al. [7] used LABView to processing voice signals which are submitted via Bluetooth to a robot whom performs movements. The use of LABView permitted to optimize the implementation costs and increase confidence in this system.

In the same context, Tereda et al. [8] developed a robot controlled by voice commands. A series of tests were performed which allowed to knowing voice commands must be improved in order to optimize recognition time. Sidiq et al. [9] presented an Android-based application named Vomma. This application implements voice recognition and execute installed applications on the same device. A set of tests was performed in order to verify Vomma works well. As salient conclusions, factors such as environmental noise, the distance between microphone and user, microphone angle and pronunciation are essential to the proper functioning of voice recognition. Costa et al. [10] used an MDD (Model-Driven Development) for an investigation of user interface named UI Stereotype (User Interface) in order to optimize the development process of Web sites.

Table 1. Comparative table between the related works with the present work.

Article	A	B	C	D	E
Erić et al. [1]	✓	N/A	N/A	N/A	N/A
Cortes et al. [2]	✓	✓	✓	✓	N/S
Modak et al. [4]	✓	✓	N/A	N/S	N/S
Tereda et al. [8]	✓	N/A	N/A	N/A	N/A
Sidiq et al. [9]	✓	N/A	✓	N/A	N/A
Costa et al. [10]	N/A	✓	N/A	✓	N/S
Paschou et al. [12]	✓	✓	✓	✓	✓
Tabibian. [13]	✓	N/A	N/A	N/A	N/A
Furtado et al. [14]	✓	N/A	N/A	N/A	N/A
Sánchez et al. [15]	✓	✓	✓	✓	N/S

A) Voice recognition implementation, **B)** Web-oriented, **C)** Mobile-oriented, **D)** Using user interface patterns, **E)** Health-oriented
N/S: Not Specified
N/A: Not Available

In addition, UI Stereotype was used to build different Web sites with the same purpose (look and feel design and functionality) from a common set of meta-models. UI Stereotype will be implemented for IFML (Interaction Flow Modelling Language) as future work. In [11], a work scheme for developing of software for processing and analyse images based on thesaurus tables' generation was proposed. Paschou et al. [12] developed a generator of health-oriented Android applications whom is used as a Web application and technical knowledge is not required to use it. Tests were performed in order to measure time, human effort and prevent constant development errors, to mention but a few. Tabibian [13] suggested a voice recognition framework in the aerospace area. This framework recognized voice commands in an acceptable time. Nevertheless, the framework functionality is affected by some factors such as pronunciation and environmental noise.

Furtado et al. [14] used Coruja Software to perform a series of tests of software in order to measure user interface performance and interaction with voice recognition. These tests allowed to know that performing some activities with voice recognition is difficult for users. According to the performance time, they concluded that the mouse is efficient while voice recognition is effective. Sánchez et al. [15] developed a software component that generates user interfaces of mobile applications by using pattern recognition and neuronal networks. A use case was proposed in order to generate a login for an Android-based application. As a result, the generation of the application was successful, however, it is necessary to perform further tests and improve patterns design. Also, the generation of user interfaces from a picture or a *mockup* will be included as future work.

Table 1 shows a comparative table among related works and this project, the table indicates the domain for which the work is focused.

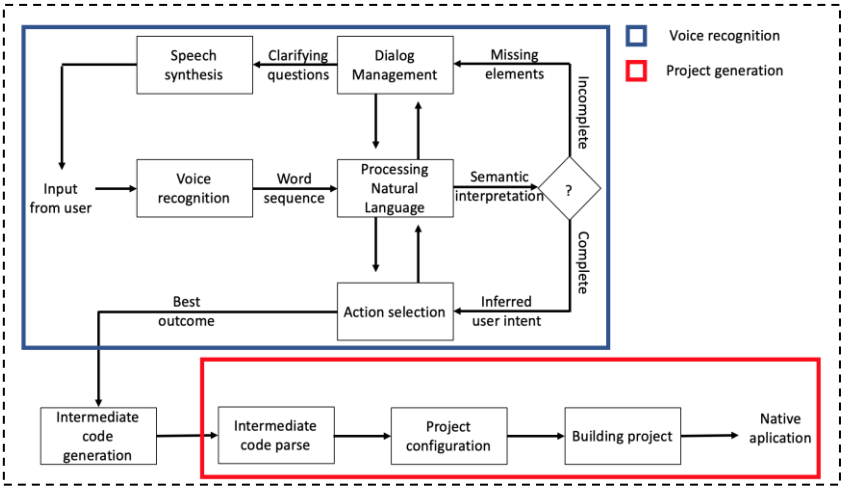


Fig. 1. Process description.

As can be seen in Table 1, the increase of development of voice recognition technologies encourages to develop efficient tools focused on different research areas. Nevertheless, it is important to considerate factors such as 1) environmental noise, 2) microphone quality, 3) pronunciation, 4) distance and angle between microphone and user, among others. Most of these research works conclude that aforementioned factors affect to voice recognition's performance. Moreover, aforementioned works related presented useful techniques, methods and technologies of voice recognition. Conversely, mostly of the related work is not focused on automatic code generation. In conclusion, none of the related works analyzed proposes a formal process for the automatic generation of mobile applications by using voice recognition. Moreover, this analysis allowed to identify a process for code generation and to considerate important factors for developing a Web-based prototype by implementing voice recognition.

3 Description of the Process for Automatic Generation of Medical User Interfaces

The process proposed in this work is divided into two phases which consist in a set of steps where different programming languages, frameworks and voice recognition technologies can be implemented. The interaction between these phases is described below in Fig. 1.

1. **Voice Recognition.** In this phase user's speech is recorded and save it as an audio file in order to turn it into word sequence which is analysed to detect user intention. NLP (Natural Language Processing) and Semantic technologies are used to identify elements of user intention such as application type, layout, user interface design patterns, platform, and information of user and application. When user intention is not completed, the dialog management module ask user

for missing elements that help to clarify user intention. For this purpose, dialog management module uses speech synthesis technologies to improve user interaction. This process is iterative until user intention is completed i.e., when all aforementioned elements were identified. Once user intention is completed an XML-based document is generated. This document describes features of application to be generated by project generation phase.

2. **Project Generation.** phase analyses the XML-based document generated in the previous phase. Then, project configuration module selects all elements that will be integrated in a software application. Finally, source code of this application is generated.

The main modules of the phases are described below with more details:

- **Voice recognition**: In this module, speech is processed to analyse word sequence and turn into text. This text will be processed to determinate its semantic.
- **Natural Language Processing**: Semantic text is determined in this module to detect user intentions. Whether user intention is completed then it will be submitted to action selection module. Otherwise dialog management module is notified to ask user for missing elements by using speech synthesis technology.
- **Action selection**: This module selects action according to user intention. An XML-based document is generated to describe all features of software application. This document is submitted to phase of project generation to generate the software application.
- **Dialog management**: The goal of this module is to ask user for missing element when NLP module does not have all elements to detect user intention.
- **Speech synthesis**: The main function of this module is to convert questions as a text file from dialog management module into an audio file. Then this module plays audio file to make a question or give a result to user.
- **Intermediate code generation**: In this module semantic and syntactic validation of XML-based document is performed to configure the final project.
- **Project configuration**: The project's features are defined in this module. This configuration depends on XML-based document's description.
- **Project building**: Finally, in this module the software application is built according to specified features through the process proposed.

4 Prototype using Voice Recognition

As a proof-of-concept, a Web-based application prototype with voice recognition was developed to segment the process. This prototype helps to have a better idea for developing an agent conversational that will provide a better user experience.

In order to develop this prototype, a workflow was organized into seven steps: 1) Selection of application type, 2) Specification of user interface layout, 3) Selection of patterns design, 4) Specification of platforms, 5) Description of user and application

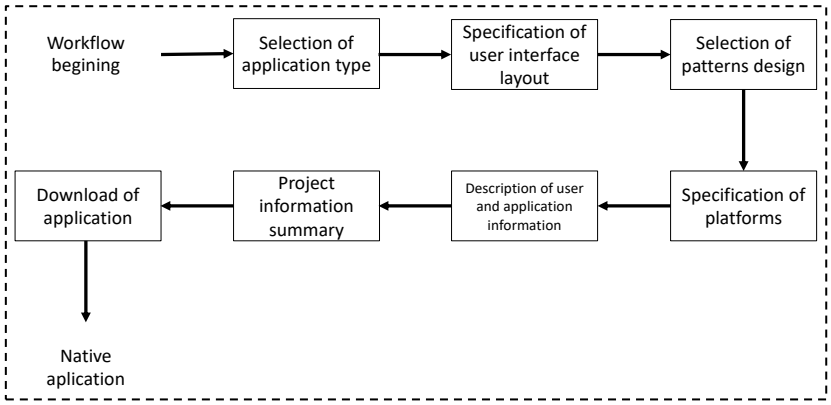


Fig. 2. Prototype's workflow.

information, 6) Project information summary, finally, 7) Download of application, as can be seen in Fig. 2.

The workflow is guided by a virtual assistant that works with voice recognition to improve user performance.

Fig. 3A depicts the first step. In this step, available medical applications are displayed; virtual assistant asks the user for the application type. Depending the context of use and kind of user, there are two types of medical apps: 1) apps for health professionals and 2) apps for the citizens. The user selects an application type by using voice recognition then a medical application is selected automatically. Afterward, a layout of application elements must be specified, a carrousel of different layout templates is displayed as can be seen in Fig. 3B.

After the selection of layout template, a set of user interfaces design patterns according to the type of application is displayed. The most used user interface design patterns in medical apps are, Datalist, Login, Gallery, Video, Splashscreen, Map, Form and Menu, to name a few. Each UIDP (User Interface Design Pattern) is used according to the characteristics and functionality that the application has.

For example, in applications to locate medical services the most used UIDPs are the Maps and Datalist. The Maps has support to the geolocation and the Datalist allows listing the health services units. Another example is the use of Search, Gallery and Dashboard patterns in health encyclopaedias apps, where Search allows searching information about a particular pathology or medication, the gallery shows sets of images, and Dashboard presents the corresponding information.

Next, in Fig. 3C is depicted as the selection of platforms for the software application. User has to use voice commands to specify the platforms. As it can be seen in Fig. 3D information about user and software is specified e.g., author name, company name, author e-mail and author's web site, application's title, application's short name, and application's version.

Before proceeding project generation, the prototype displayed a modal with a summary of the application's information. Finally, the building process of application

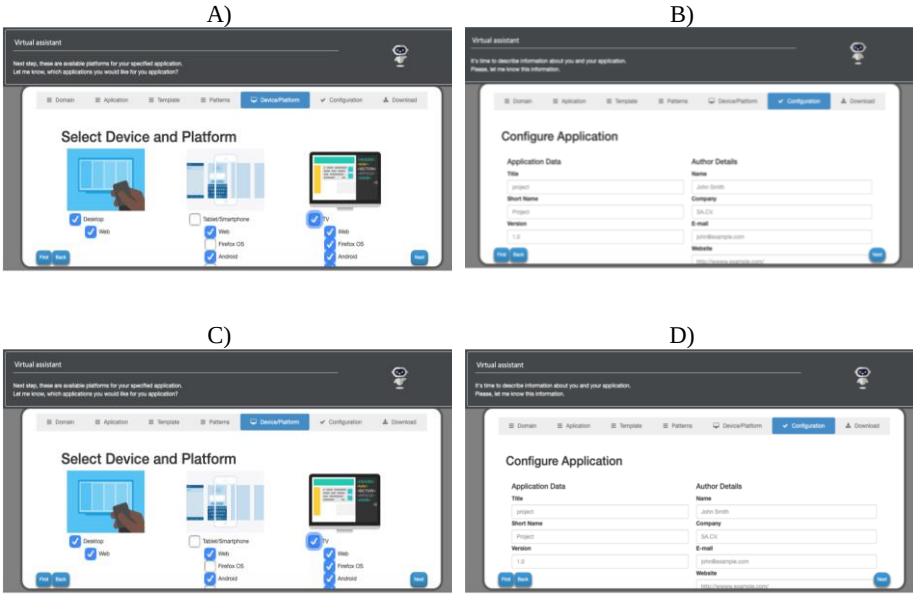


Fig. 3. Web application prototype: A) Selection of application type; B) Selection of layout template; C) Selection of devices and platforms; D) Specification of user and application information.

is executed. When the building process has finished, the software project is downloaded automatically.

5 Conclusion

The software development process has iterative activities that can be automated to optimize time of development, thus, involved people will invest time and effort in activities such as business rules, analysis, and design. Voice recognition is an innovative user interface to have a better interaction with users.

Looking for methods and techniques to improve the software development process, this paper suggests a process support by a workflow in order to generate native projects of different platforms by using voice recognition. This approach applies engineering software techniques and natural language processing which permits to hide some technical aspects to final users by a friendly user interface. Furthermore, medical applications have taken on great importance today, especially in healthcare. Due this, it is important to investigate the use of these applications in order to develop appropriate graphical user interfaces for this domain.

As future work, we are considering to develop an agent conversational to improve user experience. Likewise, research will intent to study additional user interface design

patterns and increase compatibility with other devices to promote application development in the field of healthcare.

Finally, we are evaluating to include more medical applications that allow generating other user interfaces design patterns in the medical domain.

Acknowledgments. This work was supported by Tecnológico Nacional de Mexico (TecNM) and sponsored by the National Council of Science and Technology (CONACYT) and the Secretariat of Public Education (SEP) through PRODEP (*Programa para el Desarrollo Profesional Docente*, for its acronym in Spanish).

References

1. Erić, T., Ivanović, S., Milivojša, S., Matić, M., Smiljković, N.: Voice control for smart home automation: Evaluation of approaches and possible architectures. In: IEEE 7th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), pp. 140–142 (2017)
2. Cortes-Camarillo, C.A., Rosales-Morales, V.Y., Sanchez-Morales, L.N., Alor-Hernández, G., Rodríguez-Mazahua, L.: Atila: A UIDPs-based educational application generator for mobile devices. In: International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 1–7 (2017)
3. Park Joon Young, Jo Hyo Jin, Samuel Woo, Dong Hoon Lee: BadVoice: Soundless voice-control replay attack on modern smartphones. In: Eighth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 882–887 (2016)
4. Modak, S., Vikmani, S., Shah, S., Kurup, L., Voice driven dynamic generation of webpages. In International Conference on Computing Communication Control and automation (ICCUBEA), pp. 1–4 (2016)
5. Cuello, J., Vittone, J.: Diseñando apps para móviles. José Vittone — Javier Cuello, (2013)
6. Beltramelli, T.: Pix2Code: Generating Code from a Graphical User Interface Screenshot. In: Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems, 3 pp. 1–6 (2018)
7. Patil, S., Abhigna, A., Arpitha, Deepthi, Priyanka: Voice Controlled Robot Using Labview. In: International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C), pp. 80–83 (2018)
8. Terada, H., Makino, K., Nishizaki, H., Yanase, E., Suzuki, T., Tanzawa, T.: Positioning Control of a Micro Manipulation Robot Based on Voice Command Recognition for the Microscopic Cell Operation. In: Advances in Mechanism Design II, pp. 73–79 (2017)
9. Sidiq, M., Budi, W.T.A., Sa'adah, S.: Vomma: Android application launcher using voice command. In: 3rd International Conference on Information and Communication Technology (ICoICT), pp. 49–53 (2015)
10. Costa, S.L.d, Neto, V.V.G., Oliveira, J.L.d: A User Interface Stereotype to Build Web Portals. In: 9th Latin American Web Congress, pp. 10–18 (2014)
11. Nedzved, A., Gurevich, I., Trusova, Y., Ablameyko, S.: Software development technology with automatic configuration to classes of image processing problems. Pattern Recognition and Image Analysis, 23(2), pp. 269–277 (2013)
12. Paschou, M., Sakkopoulos, E., Tsakalidis, A.: EasyHealthApps: e-Health Apps Dynamic Generation for Smartphones & Tablets. Journal of Medical Systems, 37(3), pp. 9951 (2013)

13. Tabibian, S.: A voice command detection system for aerospace applications. International Journal Speech Technology, 20(4), pp. 1049–1061, (2017)
14. Furtado, L., Marques, A., Neto, N., Mota, M., Meiguins, B.: IVOrpheus 2.0 - A Proposal for Interaction by Voice Command-Control in Three Dimensional Environments of Information Visualization. In: Human Interface and the Management of Information: Information, Design and Interaction, pp. 347–360 (2016)
15. Sánchez-Morales, L.N., Alor-Hernández, G., Miranda-Luna, R., Rosales-Morales, V.Y., Cortes-Camarillo, C.A.: Generation of User Interfaces for Mobile Applications Using Neuronal Networks. In: New Perspectives on Applied Industrial Tools and Techniques, J. L. García-Alcaraz, G. Alor-Hernández, A. A. Maldonado-Macías, and C. Sánchez-Ramírez, Eds. Cham: Springer International Publishing, pp. 211–231 (2018)

ISSN 1870-4069

A Life Cycle Cost Analysis in Wind Energy Projects in Colombia

Angélica M. González O.[1,2], Cuauhtémoc Sánchez Ramirez[2], Diego Fernando Manotas Duque[1], Magno A. González Huerta[2], Yara A. Jiménez Nieto[3]

[1] Universidad del Valle,
Departamento de Ingeniería Industrial,
Colombia

[2] Instituto Tecnológico de Orizaba,
División de Investigación y Estudios de Posgrado,
Mexico

[3] Universidad Veracruzana Campus Ixtaczoquitlán,
Facultad de Contaduría y Administración,
Mexico

angelicagonzalez0512@gmail.com,csanchezr@ito-depi.edu.mx,
diego.manotas@correounivalle.edu.co, magnogh@yahoo.com.mx,
yjimenez@uv.mx

Abstract. The demand for energy has been increasing and with this the need to find new sources of electricity generation for the system. In Colombia, different investigations have been developed in which different sources of generation are considered. In this article, we will be based on the study done in de Atlas of the wind of Colombia considering wind power as a source of generation. This article presents a case study in the LCOE analysis in generation projects of renewable energy through wind technology. Crystal Ball is used for this analysis as a simulation tool for different scenarios in order to determine the cost kilowatt hour in a wind farm in Colombia. In addition, in this article evaluates the financial feasibility of installing a wind farm in Colombia by comparing the cost kilowatt hour of energy generated in different regions of the country with the energy costs presented in the last year in Colombia and contemplating the energy policies of Colombia that generate economic benefits in the installation of renewable energy generation sources.

Keywords: LCOE analysis, renewable energy, wind energy.

1 Introduction

As energy demand grows worldwide, it is growing the need to look for new generation sources such as solar, wind, biomass, among others. Over time its use has become one

of the solutions more effective and efficient for sustainable development according to a study by [5].

If generalized measures are not implemented and effective to meet the demand for energy such as efficient production, technological developments and environmental awareness, according to studies registered in [9] the demand for primary energy is likely to expand 2 to 2.5 times by 2050; And 3.5 to 4.5 times for the year 2100.

Renewable energies generate different benefits, as mentioned by the authors [12] where they state that demand growth and decline in prices they can be better than those predicted by conventional energy sources.

This paper studies the possibility of installing a wind farm in Colombia, this study is made due to the volatility in energy prices and the decrease in emissions of CO_2 generated by wind energy and a financial analysis has been taken into account. For this, the document Atlas of Colombia made is taken as a reference by [16], which shows a collection of maps where with the distribution surface wind space shows the country's wind potential.

The financial analysis is based on an LCOE model given by the ratio between the annual cost of energy produced in a year for a certain place and the density of energy generated in said place. The LCOE methodology (Levelized Cost Of Electricity) is used to compare costs kilowatt hour between different renewable energy production projects and determine the minimum value at which the energy produced can be sold. The LCOE metric is calculated with the cost ratio and energy density as shown in equation (1) that generates the cost per kilowatt hour ($\$/KWh$):

$$LCOE = \frac{CAE_t}{P_n}, \tag{1}$$

where CAE_t is the equivalent annual cost that includes: maintenance cost, cost of operation and initial cost; t is the period in years that for this work are considered 20, and P_n is the density of energy generated during a year.

2 Methodology

For this study, the work done by presented on [16] the map is taken into account of Colombia a monthly average wind and this is carried out the methodological steps. For this study, we consider:

- Places: 16 places are located in different departments of Colombia that they are classified by [16] as the places with the best wind potential in the country.
- Average speed: Places that have an average wind speed are taken greater than or equal to 3m/s for being places where a wind farm could be installed. This information is obtained in [16].
- Data: in the cost analysis life cycle was considered the cost of a wind turbine, variations that are modeled with different distributions such as the average

wind speed, wind energy density, cost per square meter of the land where it would take the installation of the wind farm and maintenance costs thereof.

2.1 Formulation

The amount of energy transmitted to the rotor is expressed as density of power, which is directly related to air density. Power density o wind energy density (*P/A*) is given by equation (2) that depends on the density of air (ρ) and wind speed (v), these are measured in kilograms per cubic meter and meters per second, respectively.

$$\frac{P}{A} = \frac{1}{2}\rho v^3 . \tag{2}$$

The density of the air is given by:

$$\rho = \frac{P}{TR^*}, \tag{3}$$

with:

$$R^* = R\left(1 + \frac{e}{P}\right), \qquad e = \exp\left(\frac{-6763{,}6}{T}\right) - 4{,}9283\ lnT + 54{,}23,$$

the universal constant of ideal gases is *R* with a value of *286; 8Jkg/K*, *P* is the pressure atmospheric expressed in Newton per square meter and *T* is the air temperature expressed in degrees Kelvin.

The table 1 shows the data obtained from equation (3) and the energy density obtains by the equation (2) for the different places.

To carry out the modeling based on the LCOE methodology, an analysis is carried out in the which are considered the following places: Camilo Daza airport, Galerazamba, El Embrujo airport, Sesquicentenario airport, and Gachaneca. These places were determined as those with a lower coefficient of variation in wind energy density. Subsequently, they are collected data to calculate the LCOE in each of these places, a characteristic that they share these places is the wind turbine that would be installed, so a quote is made for a wind turbine through the Aelos wind turbine company.

The cost per square meter varies by location, so a price per meter is made square as close as possible to the possible location of the wind farm, due to the variety that presents the cost per square meter within the same place, in each case models a triangular distribution where the minimum, maximum and maximum prices are considered most likely as found in different sources of real estate sales.

According to different wind power generation projects, maintenance costs vary as the project progresses, starting with 2% on the cost of the investment and ending at the end of the useful life of a wind power generation project that considered twenty years in with 3% on the same cost. At the moment it is not known exactly in some reference the moment in which this change can be determined, by this reason the triangular distribution is used to model this variable, considering as value minimum 2%, the most likely 2.5% and the maximum as 3% of the initial investment.

Table 1. Matrix of data used for analysis.

Place	Average wind speed (m/s)	Average air density (kg/m^3)	Average wind energy density (W/m^3)
Galerazamba	5,9	1,15	119,19
Gachaneca	5,5	0,89	74,14
Sesquicentenario airport	5,1	1,15	76,35
La Legiosa	4,1	0,96	33,16
El embrujo airport	4,0	1,12	35,98
Almirante Padilla airport.	4,0	1,14	36,67
Obonuco	3,5	0,86	18,53
Camilo Daza airport	3,3	0,89	16,07
Urrao	3,0	0,96	17,72

Table 2. LCOE values without tax benefits.

Place	Minimum	Maximum	Most probable
Galerazamba	641,81	685,41	662,72
Sesquicentenario airport	961,34	1047,46	1004,51
Gachaneca	1011,89	1078,33	1045,78
El embrujo airport	2135,99	2250,45	2190,58
Camilo Daza airport	4517,48	4867,67	4682,36

Table 3. Correlation between the LCOE without tax benefits and parameters.

Place	Land purchase	Maintenance cost	Wind energy density	Average speed of energy
Galerazamba	-0,4983	-0,805	-0,2666	0,0029
Sesquicentenario airport	-0,7148	-0,6281	-0,2091	0,0045
Gachaneca	-0,4833	-0,8271	-0,2351	-0,0133
El embrujo airport	-0,3088	-0,9156	-0,2343	-0,0099
Camilo Daza airport	-0,6757	-0,6829	-0,2215	-0,0149

On the other hand, the average wind speed is taken as a variable for the LCOE because to the variation that it presents month by month for the same place. This is modeled with a distribution Weibull with scale and shape parameters that differ for each month, the parameters of this distribution are taken from the work done by the [16].

Because the density of the energy given by equation (2) varies are the wind speed, energy density is a variable that changes for each month, which is why taking account that its minimum, maximum and most probable value is known, and that the amount of sample data is limited, this variable is modeled with a triangular distribution taking as parameters the minimum and maximum value of the energy density for the year, and as more value probable energy density for each month.

Taking into account the variations and distributions mentioned above, it is done the financial analysis calculating an LCOE for each month in each place. For these calculations you use the Crystal Ball application and make a simulation with 5000 iterations taking as parameters land cost per square meter, maintenance cost, speed average wind and energy density. The LCOE is defined as a dependent variable.

3 Results

The results of the simulations performed with the application Crystal Ball for the calculation of the LCOE in each place. Simulations include two cases: without tax benefits and with tax benefits. These benefits include a deduction special income tax and accelerated depreciation that are granted by law 1715 of the year 2014 in Colombia. This section first presents the descriptive analysis of the sample and the latent variables. Then, we discuss results from the model evaluation, including its effects.

3.1 Without Tax Benefits

Table 2 shows the results of the minimum, maximum and most probable value of the LCOE value for each of the places. These are ordered ascending according to the LCOE obtained. In this you can see that the place with the lowest cost of sale is Galerazamba and the highest cost is from Camilo Daza airport, meaning the LCOE of Galerazamba about 14% of the most likely value of the LCOE in the Camilo Daza airport.

Table 3 shows the correlation values between the parameters taken for the simulation and the LCOE of each place.

This shows that the highest correlation in the most places are presented in the cost of maintenance and are negative, which means that for each place the parameter that most negatively affects the LCOE is the cost of maintenance, that is, the higher the maintenance cost, the higher the LCOE. On the other hand, at the Sesquicentenario and Camilo Daza airports the correlation is high with respect to the purchase of the land, this is caused by the high costs per square meter of the land with regarding the other places.

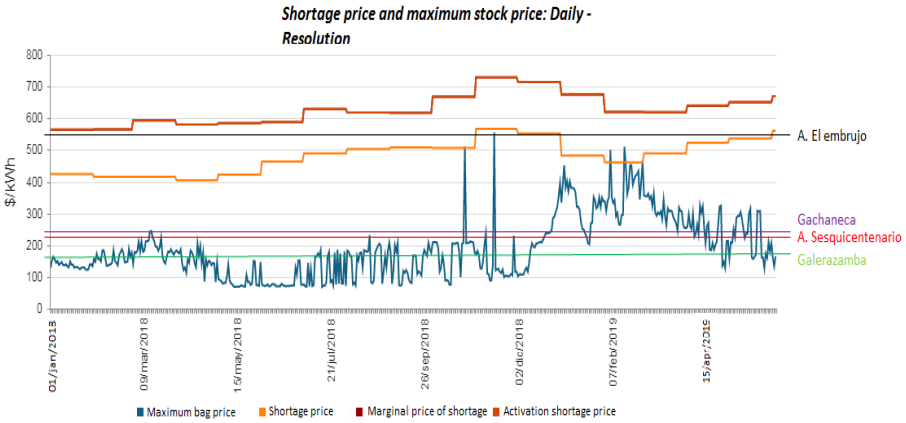


Fig. 1. Stock price compared to calculated LCOE values.

Table 4. LCOE values with tax benefits.

Place	Minimum	Maximum	Most probable
Galerazamba	148,37	189,33	168,69
Sesquicentenario airport	195,99	276,33	237,65
Gachaneca	226,77	289,15	258,72
El embrujo airport	497,12	615,89	554,97
Camilo Daza airport	874,90	1242,57	1056,14

Table 5. Correlation between the LCOE with tax benefits and parameters.

Place	Land purchase	Maintenance cost	Wind energy density	Average speed of energy
Galerazamba	-0,3766	-0,9031	-0,0655	0,0008
Sesquicentenario airport	-0,4351	-0,8888	-0,0774	0,0029
Gachaneca	-0,3658	-0,9167	-0,0562	-0,0188
El embrujo airport	-0,2121	-0,9716	-0,0564	-0,0152
Camilo Daza airport	-0,5482	-0,8149	-0,0377	-0,0083

3.2 With Tax Benefits

As of 2014, tax benefits are generated according to Law 1715 "Por medio de la cual se regula la integración de las energías renovables no convencionales al Sistema

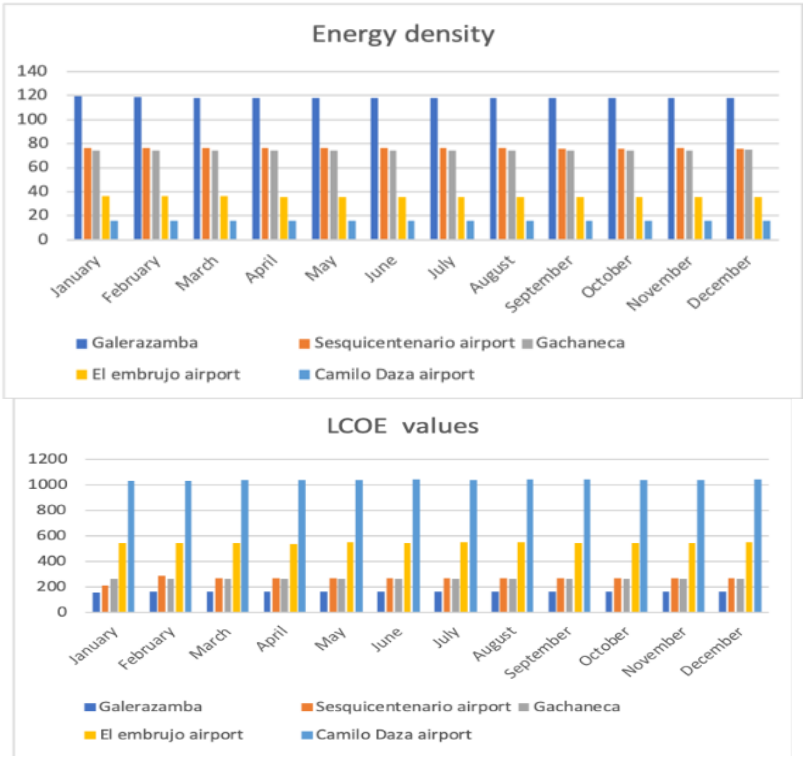


Fig. 2. Graph of LCOE relationship and energy density.

Energético Nacional". Given this law, for this project is considered the special deduction of income tax that consists of deducting for each taxable year a value not exceeding 50% of the taxpayer's liquid income, before to subtract the deduction; and accelerated depreciation where the global annual depreciation rate. It cannot exceed 20%. This generates changes in the LCOE for each place, which is why run the simulations for each zone again.

Table 4 shows the results of the LCOE analysis with tax benefits for Each of the places. These are sorted in ascending order according to the LCOE obtained. It can be seen that the place with the lowest cost of sale is Galerazamba and the one with the highest cost is Camilo Daza airport, the LCOE of Galerazamba being around 16% of the value plus LCOE likely at Camilo Daza airport.

Table 5 shows the correlation values between the parameters taken for the simulation and the LCOE with tax benefits of each place. This shows that the higher correlation is presented with the cost of maintenance and is also negative, which means that for each place the parameter that most negatively affects the LCOE is the cost of maintenance. That is, regardless of where the wind farm is installed, a decrease in the cost of maintenance would generate a decrease in the LCOE. Sales costs calculated in the analysis take into account the benefits tax and are similar to the costs obtained in the financial analysis made by the authors in [14].

This analysis is done with seven different power turbines between 6,000 kW and 2,750 kW in regions of Colombia that are not connected to the national electrical system and have an annual average speed greater than 4m/s. This shows that variations in the financial analysis of these types of projects. It does not affect the cost of sale kilowatt hour. Therefore, there are different variables that are they can take into account for the execution of a wind power generation project, giving relevance to social, environmental and economic aspects.

Image 1 shows the stock and shortage prices given by XM (www.xm.com.co) on a daily basis from January 1, 2018 to June 4, 2019, in addition, LCOE average values were added considering the tax benefits granted in law 1715 that were applied to each of the five places studied in this document, these values behave similarly in each month as observed in the different overlay graphics shown in this section.

The Energy Density chart consolidates the energy densities for the five places with lower coefficient of variation and on the LCOE chart considering benefits tax values are consolidated in the base case of the statistics for each place.

When comparing these two graphs, it can be seen that the higher the density value of wind energy for each place in a general way, the lower the LCOE corresponding to the same place. Preserving this correspondence for each place, which could be assumed that one of the main factors that influence the LCOE corresponding to each place is the density of energy present in said place. This observation is valid taking into account the denominator of the formula equation 1.

4 Conclusions

The key factors for the installation of a wind farm are the average speed of the wind and wind energy density, these are used for the life cycle cost analysis. From the average speed 16 places are selected with wind potential in Colombia and with the density of wind energy the production of wind energy in different places.

The generation capacities of the wind farms for the five locations in order to coefficient of variation of wind energy, and determined, from the LCOE analysis, that Galerazamba and Gachaceca are the best places for the installation of a wind farm in Colombia.

The maintenance cost is a determining factor to calculate the LCOE in each place and the tax benefits of Law 1715 of 2014 are a great incentive for the execution of a wind farm in Colombia, reducing the LCOE between 74% and 77% in each place.

The variation that occurs in wind energy density per month and the variation in Weibull distribution parameters that vary in average speed for each month, are parameters that modify the output variable, the smoothed cost of LCOE energy per month in each of the places.

When considering the tax benefits granted in law 1715 of the year 2014 in this class of technology makes it competitive against the market prices observed in the market wholesaler for Galerazamba, A. Sesquicentenario, Gacaneca and A. El embrujo.

References

1. Akella, A.K., Saini, R.P., Sharma, M.P.: Social, economical and environmental impacts of renewable energy systems. Renewable Energy, 34(2), pp. 390−396 (2009)
2. AccuWeather: https://www.accuweather.com (2017)
3. Mapcarta: https://www.mapcarta.com (2017)
4. Díaz, M.A.R.B. Energía eólica. Boletín IIE (2013)
5. Dincer, I.: Renewable energy and sustainable development: a crucial review. Renewable and Sustainable Energy Reviews, 4(2), pp. 157−175 (2000)
6. Escudero, J.M., López, J.M.E.: Manual de energía eólica/guide to wind energy. Mundi-Prensa Libros (2008)
7. Gálvez, R.: Diseño y cálculo preliminar de la torre de un aerogenerador (2005)
8. Grubler, A., McDonald, A.: Global energy perspectives. Nakicenovic, Global Energy Perspectives (1998)
9. Guillen, S.O.: Energía eólica para generación eléctrica. Trillas (2015)
10. Herzog, A.V., Lipman, T.E., Kammen, D.M.: Renewable energy sources. Encyclopedia of Life Support Systems (EOLSS), Forerunner, Perspectives and overview of life support systems and sustainable development (2001)
11. IRENA: Renewable power generation costs in 2017 (2017)
12. Johansson, T.B., Burnham, L.: Renewable energy: sources for fuels and electricity. Island Press (1993)
13. Johnson, G.L. Wind energy systems englewood cliffs (NJ). Prentice Hall, pp. 147−149 (1985)
14. Jimenez, R., Diazgranados, A., Acevedo-Morantes, M.T.: Electricity generation and wind potential assessment in regions of Colombia. Dyna, 79(171), pp. 116-−122 (2012)
15. San Cristóbal, M.J.R.: Multi-criteria analysis in the renewable energy industry. Springer (2012)
16. UPME: Atlas de viento y de energía eólica de Colombia (2006)

ISSN 1870-4069

A Multi-Agent System for the Inventory and Routing Assignment

Conrado Augusto Serna Urán[1], Cristian Giovanny Gómez Marín[2], Julián Andrés Zapata Cortes[3], Martín Darío Arango Serna[2]

[1] Universidad de San Buenaventura, Colombia

[2] Universidad Nacional de Colombia, Colombia

[3] Institución Universitaria CEIPA, Colombia

conrado.serna@usbmed.edu.co

Abstract. In the supply chain management research field, the analysis strategies and the joint management of inventory and transport of goods have increased attention because of current changes and tendencies in the goods interchange process. However, the computational complexity and the problems that may arise in the integration processes of the different actors become the majority of proposals difficult to implement. In this paper, we develop a multi-agent system (MAS) for solving the inventory and routing assignment problem. The proposed multi-agent system facilitates the integration of the distribution processes and the inventory management in a supply network with one depot and n customers. The model is based in the autonomy of the actors to manage their capacity, demand, and integration of the transport process. To solve the resulting vehicle routing problem, we use a 2-opt local search heuristic.

Keywords: inventory routing problem, multi-agent system, supply chain management, supply chain collaboration.

1 Introduction

The integration of the distribution and inventory process is one of the main strategies to produce cost efficiency and satisfactory service levels among the actors of the supply chain. The objective of this integration is to reduce the global cost of the distribution process, which include the holding and the transport costs (Zapata, 2016).

The implementation of this type strategy is difficult due to two different problems: the mathematical complexity of the problem and the fact that the involved actors try to conserve specific and individual objectives.

As a solution to these difficulties, some authors have found in the multi-agent systems a promising alternative because of its versatility and ease implementation. The multi-agent systems use the distributed computing paradigm (Arango-Serna, Serna-Urán, & Zapata-Cortes., 2018), furthermore, the use of multi-agent systems facilitate the representation of the supply chain actors in artificial agents with their capacities and specific goals interacting in a coordinated and collaborative function system. The multi-agent systems consist in multiple agents interacting to solve a common problem, compete for the use of shared resources or simply they coordinate each other to avoid conflicts (Arango-Serna & Serna-Urán, 2016; Serna-Urán, 2016).

This article presents a Multi Agent simulation-based model for the freight distribution process. The model seeks to reduce the total distribution costs involving the decision of the transportation and inventory plans. In the first part of the article, a short literature background about the integration of transportation and inventory decision, and Multi agent Systems is presented. Then, the propose multi agent model is explained and later applied in a median size problem. Finally, the analysis and some conclusions are stated.

2 Background

Guerrero et al. (2013) argue that inventory management and distribution decisions are related to each other, since inventory depends on the frequency and time to supply companies, as well as orders and product costs (Arango-Serna et al., 2015). The decision of simultaneously assign the inventory and transportation is carried out following two types of mechanisms: Decomposition and aggregation: The decomposition mechanism separates the problem into two phases, in which the first determines the inventory and the second the transportation routes (Kang and Kim, 2010).

The aggregation mechanism finds the solution to the problem simultaneously, producing the inventory and transport decisions directly. To solve this problem, the most studied model is the Inventory Routing Problem (IRP) (Arango-Serna et al., 2016), which is based on the Vendor Managed Inventory and seeks to combine the transport and inventory allocation problem to reduce the global distribution costs (Arango-Serna et al., 2016).

In recent years, the supply chain management have received much interest to be modeled as a multi-agent system (Aminzadegan, Tamannaei, & Rasti-Barzoki, 2019; Avci & Selim, 2016; Ghadimi, Ghassemi, & Heavey, 2017; Goncalo & Morais, 2016; Kumari, Singh, Mishra, & Garza-reyes, 2015; Pal & Karakostas, 2014; Sitek, Wikarek, & Grzybowska, 2014), in which each agent seeks to maximize his profits instead of the general benefit of the supply chain. Each agent acts autonomously, interacts with other agents, reacts to changes in the environment and makes proactive decisions (Böhnlein, Schweiger, & Tuma, 2011; Wang, Wong, & Wang, 2013).

Table 1. Agents description. own source.

Agent name	Description
Customer Agent	It asks for service order defined by quantity, location and period of time of their demand. The agent customers base their decision process on the inventory cost analysis criterion. They expect to be served in a just in time process.
Service Control Agent:	Builds delivery services defining the quantity, period and location of the delivery. Search for reducing the transportation cost and augmenting the use of the vehicles without exceed their capacity form the cluster of requests.
Route Agent	Builds the routes of services using the 2-OPT local search heuristic.
Collaboration (integrator) Agent	This agent explores solutions for transportation cost reduction. The customers with a high ratio between holding and transportation costs are proposed to integrate other service routes. The transportation cost is computing from the cost of insert the customer to an evaluated route.

In this sense, the decision process of integrate the minimization of inventory and transportation costs can be modeled as a multi-agent system, in which customers, suppliers and transporters interact. Regardless every actor has their own interests, they must be balanced to optimize the total costs of the supply chain. This modelling is presented in the next section.

3 Multi-Agent System for the Inventory and Routing Decision

Taking advantage of the benefits of distributed computing, the decentralization of tasks and the coordination and integration between supply chain actors, a multi-agent system (MAS) to solve the combined problem of inventory and routing assignment was developed.

This multi-agent system is based on the representation of the customers and the transport-logistics operator as virtual agents that reproduce the input data from the physical system to the virtual system. The logistic operator develops the information exchange and the coordination processes among all agents in the multi-agent system. The multi-agent model is composed by the agents presented in table 1.

By using a *collaborative strategy*, the logistic operator (*Agent service control*) requests the customers (*Agents customer*) for the demand information at the *t* period, in order to generate the service orders. The customers may accept or reject the proposal of the *Agent service control*. If the proposal is accepted, the Agent service control request to the agent route to build a route for each service at the t period. Finally, Agent Collaboration searches for possible improvements by changing requests on the routes based on the location of the client and the period *t* of the demand.

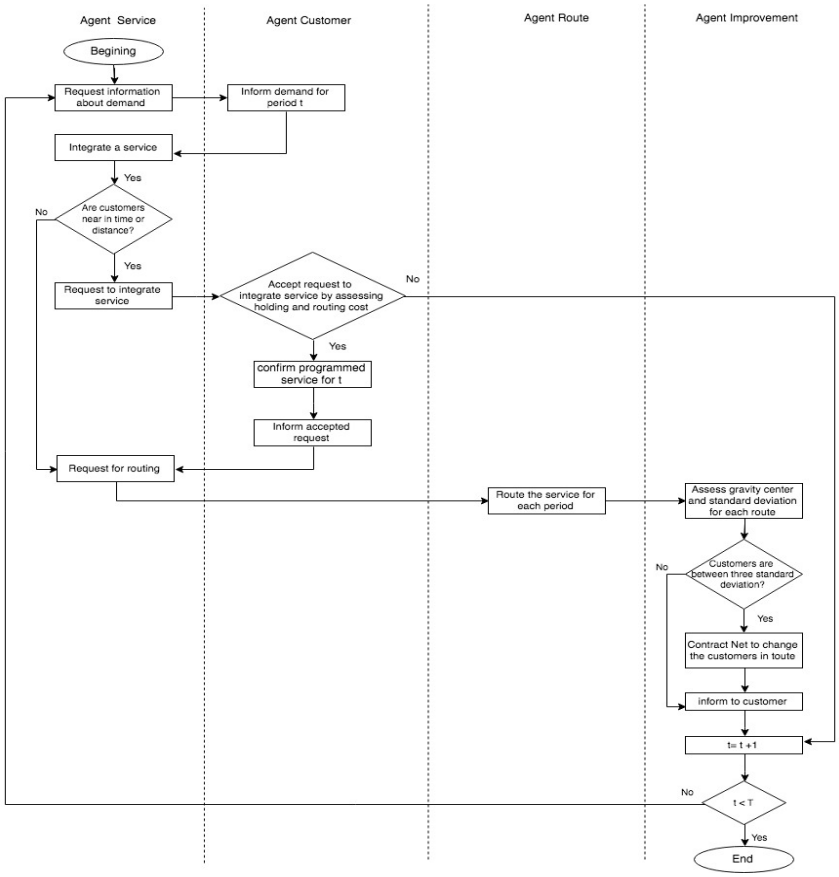


Fig. 1. Communication process at the multi-agent system. Own source.

The communication and coordination process carried out by the model can be depicted in Figure 1, and it is explained below:

The communication protocols between agents allow generating flexibility in the multi-agent system, through the structured management of the information that each agent receives and sends. In the multi-agent system designed, the interaction between the agents is ruled by the Request and Contract Net protocols, as can be seen in figure 2. The operations of the protocols are:

- ***Request Protocol (RP):*** The RP is a communication protocol that allows the agents to make information requests and actions to other agents with specific behaviors, capabilities and resources. This protocol allows the multi-agent system to initialize the service offers process by the logistics operator by requesting demand information. Each customer returns its demand information in period *t.* With this information, the service agent establishes the service for the period *t.*

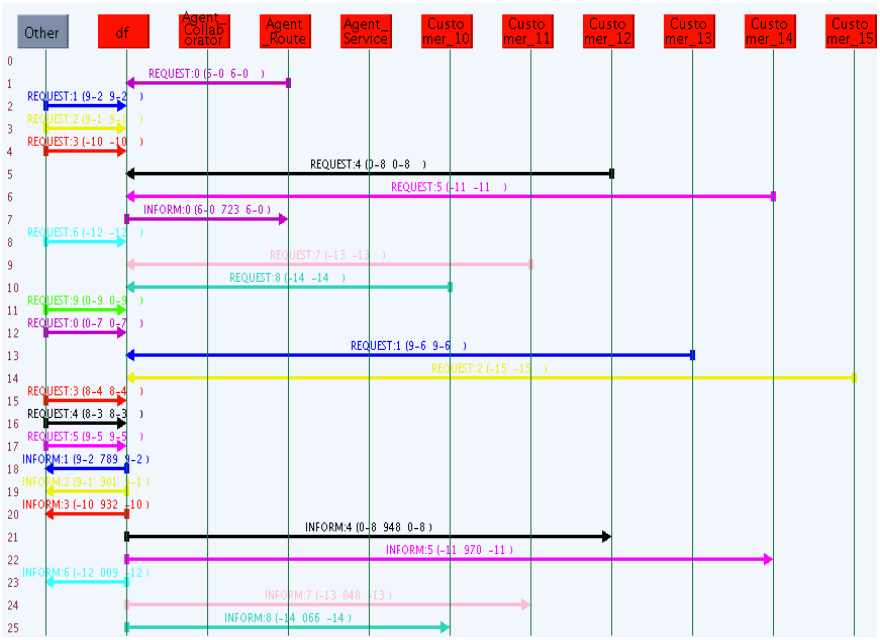


Fig. 2. Communication among agents during simulation. Own source.

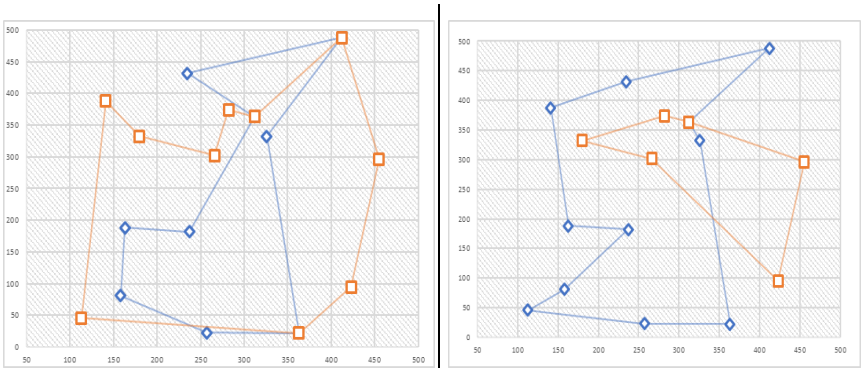
- The CNP is a negotiation protocol in which the agent that initiates the process (improvement agent) makes a request to the participating agents (agents Route) to send the cost proposals of including a customer in a service for the period $t + l$. The route agents evaluate the cost and benefit of including the customer in the existing route. With the answers of the route agents, the improvement agent selects and assigns the best proposal.

4 Multi Agent System Application

The MAS was tested using a distribution example, which consider 15 customers, one supplier and a 3-period time horizon. The simulation was carried out using JADE®. For the first period, all customers can supply their demand from the inventory, in that sense no service has been integrated by the logistic operator.

For the second period, there are two different services: the first one is integrated according with the information that the customers have send to the operator, and the second one is integrated through the interaction between customers and operator to figure out local decision for each customer. With these interactions in the second period, the customers that are subscribed for this service are supplied. Then, it is required to compute the needs for the third period.

When the service integration has finished, the agent route start to build their route using the *2-opt* local search heuristic to compute the cost of the routes for each service in each period. The results of the behavior of this agent is showed in table 2.



(a) initial solution (b) Improved solution

Fig. 3. Improvement of solution in the MAS. Own source.

Table 2. Routes for each service-period. Own source.

Period	Routes	Cost (U$)
t=2	0, 1 , 4 , 14 , 12 , 13 , 7 , 9 , 8 , 0	13,77
t=3	0, 5 , 11 , 2 , 3 , 10 , 13 , 15 , 6 , 9 , 0	15,14

Table 3. Final routes after the collaboration among agents. Own source.

Period	Routes	Transport cost	Inventory cost
t=2	0, 7 , 13 , 12 , 10 , 14 , 1 , 4 , 3 , 8 , 9 , 0	15,08	14,09
t=3	0, 5 , 2 , 11 , 15 , 6 , 0	8,54	0

With these results, the agent collaboration begins to make their analysis and trigger the contract net protocol with the agent's routes to find a better inventory and transportation cost combination for the system.

First, it should compute the gravity center, the average distance, and the standard deviation for the bigger and more expensive route. The customers that their distance to the center of gravity are bigger than three standard deviation are selected to evaluate a change in the route.

The contract net protocol asks for the routes to compute the cost of each route with and without the selected customers and the collaborator agent compute the inventory cost of make the changes in route to take the decision of change the customer of route or not. The final integrated services for periods t=2 and t=3 and their routes are presented in table 3.

With the agent collaborator, the total cost decrease US $5,89, equivalent to 15,64% less than the initial solution. Figure 3 presents the initial and improved solutions of the routes obtained with the Multi-agent system.

5 Conclusions

In this article, the combined inventory and transportation decision process in the supply chain was modeled as a multi-agent system that facilitates the analysis of the relationships between customers and suppliers in a framework of autonomy and collaboration. The model proposes an easy-to-implement and functional architecture in companies in which the inventory administration is not completely centralized. For this, different agents were designed to support the distribution and storage processes, which integrate functions and interaction protocols to make decisions based on their own resources and interests. However, the agents also evaluate their decisions based on principles of collaboration and global efficiency.

The results found by the multi-agent system do not correspond to the optimal solution, which can be obtained using the Inventory Routing Problem (IRP). Unlike the IRP, the proposed multi-agent system is more flexible, which considers the individual capacities and goals of the different agents involved in the distribution process. The multi-agent system seeks to improve the solutions in a collaborative framework through heuristic procedures that include the negotiation processes (request and contract net protocols) and the analysis of the storage conditions and the route cost.

As future research work, it can be analyzed more complex instances of the distribution process and its results should be compared with the optimal solutions produced by optimization methods such as the IRP model, with the aim of deriving how close the results of the multi-agent model are with respect to these optimal distribution plans. In addition, as future lines of research, it is suggested the integration of new agents in the model, which allow the development of simulation process closer to real situations, including suppliers, local administrators and other transport systems such as passengers and private transportation.

References

1. Aminzadegan, S., Tamannaei, M., Rasti-Barzoki, M.: Multi-agent supply chain scheduling problem by considering resource allocation and transportation. Computers & Industrial Engineering, 137(08) (2019)
2. Arango-Serna, M.D., Andrés-Romano, C., Zapata-Cortés, J.A.: Collaborative goods distribution using the IRP model. DYNA, 83(196), pp. 204−2012 (2016)
3. Arango-Serna, M.D., Serna-Urán, C.: New contract net negotiation protocol based on fuzzy inference applied to the supply chain. Universidad, Ciencia y Tecnología, 20(81), pp. 176–187 (2016)
4. Arango-Serna, M.D., Serna-Urán, C.A., Zapata-Cortes., J.A.: Multi-agent system modeling for the coordination of processes of distribution of goods using a memetic algorithm. In: Alor-Hernández, G.-A.J., Maldonado-Macías, G.A., Sánchez-Ramírez, C. (Eds.), New perspectives on applied industrial tools and techniques. Management and Industrial Engineering, pp. 71–89, Springer (2018)

5. Arango-Serna, M.D., Zapata-Cortes, J.A., Gutierrez, D.: Modeling the Inventory Routing Problem (IRP) with multiple depots with genetic algorithms. IEEE Latin American Transactions, 13(12), pp. 3959 – 3965 (2015)
6. Archetti, C., Bertazzi, L., Laporte, G., Speranza, M.G.: A branch and cut algorithm for a vendor managed inventory routing problem. Transportation Science, 1(3), pp. 382–391 (2007)
7. Avci, M.G., Selim, H.: A multi-agent system model for supply chains with lateral preventive transshipments: Application in a multi-national automotive supply chain. Computers in Industry, 82, pp. 28–39 (2016)
8. Belykh, D.L., Botvin, G.A.: Multi-agent framework for supply chain dynamics modelling with information sharing and demand forecast. Communications in Computer and Information Science, 858, Springer International Publishing (2018)
9. Böhnlein, D., Schweiger, K., Tuma, A.: Multi-agent-based transport planning in the newspaper industry. International Journal of Production Economics, 131(1), pp. 146–157 (2011)
10. Daganzo, C.: Logistics systems analysis. Berlin Heidelberg, Springer (2005)
11. Ghadimi, P., Ghassemi, F., Heavey, C.: A multi-agent systems approach for sustainable supplier selection and order allocation in a partnership supply chain. European Journal of Operational Research, pp. 1–16 (2017)
12. Goncalo, T.E.E., Morais, D.C.: Agent-based negotiation protocol for selecting transportation providers in a retail company. In Proceedings IEEE International Conference on Systems, Man, and Cybernetics, SMC, pp. 263–267 (2016)
13. Guerrero, W.J., Prodhon, C., Velasco, N., Amaya, C.A.: Hybrid heuristic for the inventory location-routing problem with deterministic demand. Int. J. Production Economics, 146, pp. 359–370 (2013)
14. Kang, J.H., Kim, Y.D.: Coordination of inventory y transportation managements in a two-level supply chain. Int. J. Production Economics, 123, pp. 137–145 (2010)
15. Kumari, S., Singh, A., Mishra, N., Garza-Reyes, J.A.: A multi-agent architecture for outsourcing SMEs manufacturing supply chain. Robotics and Computer Integrated Manufacturing, 36, pp. 36–44 (2015)
16. Lopes, F., Coelho, H.: Bilateral negotiation in a multi-agent supply chain system. Lecture Notes in Business Information Processing, 61, pp. 195–206 (2010)
17. Pal, K., Karakostas, B.A.: Multi agent-based service framework for supply chain management. Procedia Computer Science, 32, pp. 53–60 (2014)
18. Serna-Urán, C.A.: Modelo multi-agente para problemas de recogida y entrega de mercancías con ventanas de tiempo usando un algoritmo memético con relajaciones difusas. Universidad Nacional de Colombia (2016)
19. Sitek, P., Nielsen, I.E., Wikarek, J.: A hybrid multi-agent approach to the solving supply chain problems. Procedia Computer Science, 35(C), pp. 1557–1566 (2014)
20. Sitek, P., Wikarek, J., Grzybowska, K.: A multi-agent approach to the multi-Echelon capacitated vehicle routing problem. In International Conference on Practical Applications of Agents and Multi-Agents Systems, pp. 121–132, Switzerland Springer, Cham (2014)

21. Wang, G., Wong, T.N., Wang, X.: An ontology based approach to organize multi-agent assisted supply chain negotiations. Computers and Industrial Engineering, 65(1), pp. 2–15 (2013)
22. Zapata-Cortes, J.A.: Optimización de la distribución de mercancías utilizando un modelo genético multiobjetivo de inventario colaborativo de m proveedores con n clientes. Universidad Nacional de Colombia (2016)

ISSN 1870-4069

Multi-Objective Product Allocation Model in Warehouses

Julián Andrés Zapata Cortes[1], Martín Darío Arango Serna[2],
Conrado Augusto Serna Urán[3], Luisa Fernanda Ortiz Vasquez[3]

[1] Institución Universitaria CEIPA,
Colombia

[2] Universidad Nacional de Colombia,
Colombia

[3] Universidad de San Buenaventura,
Colombia

Julian.zapata@ceipa.edu.co

Abstract. The allocation of the merchandise to the different spaces available in the warehouses is a determining factor of the cost and the time of operation in said facilities, which is why it is important to establish the optimal storage positions, seeking better conditions of profitability and service for the Business. This article presents a multi-objective mathematical model to determine the allocation of the merchandise in the different spaces available in the warehouses, which simultaneously evaluates the operating costs and the times required to carry out the storage activities. The model is solved using a genetic algorithm, which due to the multiobjective nature, allows obtaining a set of possible accommodations of the merchandise, which can be used by the warehouse manager, sure that it will use an adequate allocation according to its preference in relation to the variables analyzed, which constitute an optimal relationship between the cost and the time required for storage.

Keywords: warehousing, merchandise allocation, multi-objective model, genetic algorithm, optimization.

1 Introduction

Products allocation is one of the key activities in warehouses optimization, since the appropriate selection of the places where the merchandise must be stored is a key activity to reduce operative costs and time. It is possible due to the reduction of the distances and the efforts required moving the goods inside these facilities, which allows reducing the costs of material handling and labor hours, as well as the time required for internal operations and thus the time of order fulfillment.

In the specialized scientific literature, there are several mathematical models and information communication systems that allow to optimally assign the products on the warehouses' shelves (Tompkins et al., 2010). However, these models normally only consider the individual optimization of one of the several objectives involved in real operation. This frequently leads to the optimization of the selected objective while negatively affecting others objective functions to be optimized, which is not in accordance with the reality of the companies, in which all objectives and variables are important.

This paper presents a multiobjective optimization model with the aim of simultaneously minimizing the product handling costs and the required time to fulfill the orders in an industrial warehouse. The model helps the decision maker people to make the decisions about the products allocation, considering both critical objectives for the warehouse performance. Due to the complexity of the optimization process, a multiobjective genetic algorithm of the type NSGAII (NonDominated sorting Genetic Algorithm II) was developed to solve the allocation mathematical model.

The results produced by this genetic algorithm behave according to the model expectations, yielding not a single solution, but a set of possible solutions that simultaneously optimize both objective functions. The decision maker can select any of those solutions, based on its preferences, sure that each solution generates an optimal product allocation plan considering both functions for the warehouse.

2 Product Assignation in Warehouses

Warehouses are facilities dedicated to the storage and care of products in organizations with the purpose of create a buffer to deal with the demand and supply variations (Zapata-Cortes et al., 2019; Ballesteros et al ., 2019). Storage is responsible for the generation of lots of cost in companies (Departamento Nacional de Planeación, 2018) due to the amount of economic, financial, personal and infrastructure resources required in such facilities for both the movement and storage of the goods (Chopra and Meindl, 2013).On the other hand, storage is also responsible for the service level perceived by customers, since it affect the lead time, the orders fulfillment and the quality and conditions of the products (Frazelle and Soho, 2010). Both the cost and service level are key factors for company's success (Chopra and Meindl, 2013). The costs optimization and the adequate service level are two of the main objectives pursued by warehouses' administrators, which can be achieved through multiple initiatives such as reducing the unnecessary distances and movements, improving the use of space, equipment, labor, accessibility to all items, among others.

These initiatives can be carried out through the use of information and communication technologies and information systems (Zapata et al., 2010), the appropriate facilities design (Arango et al., 2010) and the process optimization (De Koster et al., 2007). The use of technologies such as WMS (Warehouse Management System), barcode, radio frequency identification, among others, can increase the warehouse performance, reduce costs and improve the service levels of the storage operations (Zapata et al. , 2010).

An adequate design and the use of the right equipment and technology in the handling and storage processes, also impacts positively the performance of the warehouse. (Thomkins et al., 2010). The optimization refers to the adequate programming and allocation of people, equipment and materials, which can be done through mathematical procedures to determine the necessary resources, the optimal travel distances inside the warehouse for both the movement of people and equipment, the correct product location, the adequate quantities to be stored, among other activities (De Koster et al., 2007, Thomkins et al., 2010).

One alternative to improve the costs and response time in warehouse operations is to minimize the average travel distance, which means reducing the total distance to be travel in the warehouse (De Koster et al., 2007; Arango et al., 2010). This can be done by properly selecting the places where the goods should be located, which reduces the time and cost required to reach their storage positions (Zapata et a., 2019).

Many works and models found in the scientific literature consider the study of the single cost or operating time optimization (Sanei et al., 2011; Kovács, 2011; van Wijk et al., 2013; Zapata et al., 2019), while others researches perform joint optimization processes of several objectives (Ramtin & Pazour, 2015; Quintanilla et al., 2015; Hu et al. 2012), which allow to take decisions more in line with reality, where several objectives are important for the correct warehouse performance.

Multiobjective optimization differs from the conventional single-objective optimization since it does not generate a single solution, but a set of possible optimal solutions, from which the decision maker can select any according his preferences. This set of optimal solutions is known as the Pareto Frontier, where the improvement of one objective may result in the decrease of another or more objectives (Shenfield et al., 2007).

The Pareto frontier can be obtained with the use of conventional (classical) or through heuristic techniques (López et al., 2009). Classical methods present several disadvantages in the multi-objective optimization, since they require a high number of iterations to find the Pareto frontier, a prior knowledge of the problem domain, some of these methods are sensitive to the form or continuity of the Pareto frontier (López et al., 2009) and finding a satisfactory solution becomes increasingly complex as the number of objectives increase (Fonseca & Fleming, 1995). Heuristic methods allow to face the above-mentioned problems, finding good solutions (close to the optimal) in a reasonable processing time. Authors such as González (2013) and López et al., (2009) have analyzed the most used metaheuristic techniques to solve multiobjective optimization problems, among which are the NSGA and NSGA-II- (Nondominated Sorting Genetic Algorithm), SPEA and SPEA2 (Strength Pareto Evolutionary Algorithm), PAES (Pareto Archived Evolution Strategy), Multi-Objective Variable Environment Search (MO-VNS).

3 Multiobjective Model for Product Allocation in Warehouses

The product allocation to the different positions in the warehouse can be done by optimizing the costs or time required to perform all the warehouse operations, which is

possible through the objective function presented in equation 1. This equation represents the costs minimization for the allocation problem where there are multiple products and several collecting/delivery points in the warehouse (Tompkins et al., 2010):

$$F1 = minimize \sum_{j=1}^{n} \sum_{k=1}^{q} \frac{T_j}{S_j} \sum_{i=1}^{m} p_i c_{ik} x_{jk}, \quad (1)$$

x_{jk} is the binary decision variable , which takes the value of 1 if product j is assigned to position k or zero otherwise. The parameters description in the objective function is:

- n is the number of products.
- q is the number of storage positions.
- m is the number of origin/delivery points in the warehouse.
- T_j is the number of storage trips (input-output) for product j.
- S_j is the number of storage positions required for product j.
- p_i. is the percentage of trips to go to i and return.
- c_{ik} is the cost of travel from point i to the storage location k.

The objective function F1 can be modified to optimize the time required to fulfill the orders. It can be done by replacing the costs (c_{ik}) with the time required to perform this activity (t_{ik}: time required to go from i to the storage position *k*). This generates the second objective function of the proposed model, presented in equation 2:

$$F2 = minimize \sum_{j=1}^{n} \sum_{k=1}^{q} \frac{T_j}{S_j} \sum_{i=1}^{m} p_i t_{ik} x_{jk}. \quad (2)$$

In this way, the model contemplates the costs and time for the expected travels distances for any product j from each origin point to its storage position k, which is represented by the term $p_i d_{ik} x_{jk}$.

In addition, each objective function contemplates the movements intensity in terms of the storage positions with the term $\frac{T_j}{S_j}$ required for each product J. With this, the total cost and time for the established operation time T_j. are calculated.

Those objective functions are subject to the following constraints:

$$\sum_{j=1}^{n} x_{jk} = 1 \qquad k = a, \dots, q\ , \quad (3)$$

$$\sum_{k=1}^{q} x_{jk} = S_j \qquad j = a, \dots, n. \quad (4)$$

The probability that each item j travels from point i to each position k is the same for all products. Constraint 3 ensures that product j is assigned to position k only once. That is, only a product j can be assigned to a position k. Constraint 4 indicates that the quantity of products j assigned to positions k must be equal to the storage requirement (required spaces) for product j.

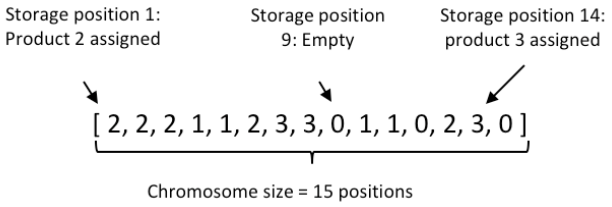


Fig. 1. Chromosome used for the product allocation problem.

4 Multiobjective Genetic Algorithm for the Product Allocation Optimization

For the solution of the multiobjective model presented above, a genetic algorithm NSGAII (NonDominated Sorting Genetic Algorithm-II) was developed. This type of algorithm is one the most used multiobjective evolutionary optimization algorithms in the scientific literature (Zapata, 2016). Unlike many evolutionary algorithms, the NSGAII has the ability to work easily with the classification of individuals according to the non-dominance criterion. It also prevents the loss of good solutions since it constantly evaluates new and old individuals and contemplates the concept of diversity among them.

In the NSGAII algorithm, an initial population $P_t = 0$ of size N is randomly created, which is organized according to the non-dominance of the individuals. This process is carried out through an iterative procedure in which different levels of non-dominance are determined, that is: Pareto Borders are obtained by separating the non-dominated solutions from the rest and then the non-dominated individuals are classified again with the non-dominance criterion. This process is carried out until all individual of the population is assigned to its non-dominance level. (Correa et al., 2008).

From this initial population, a new temporary population Q_0 of size N is generated (Offspring population), through the selection, crossover and mutation operators of the genetic algorithm. These populations are combined to form a population R_t of size 2N ($P_t + Q_t$) which is ordered according to the non-dominance levels, thus ensuring elitism in the algorithm. From the different levels of non-dominance (Pareto borders) of R_t, a new population P_{t+1} is created including (accommodating) the individuals at the best levels of non-dominance (First Pareto borders) (Deb et al., 2002; Zapata, 2019).

Every individual in the NSGAII algorithm are represented as a vector of integers numbers (chromosome), in which each position i represents a storage position, as shown in Figure 1. To each position i, an integer number from zero to the number of different types of products is assigned, so the sum of positions with the same number is equal to the number of spaces in the warehouse required to storage such product. In this representation, a value 0 means that this space is empty.

The selection of individuals is made by tournament with size m (m = 5% of the population size) which are randomly selected. These individuals are compared and selected according to the non-dominance and the dispersion criterion (Crowing Distance) (Correa et al., 2008).

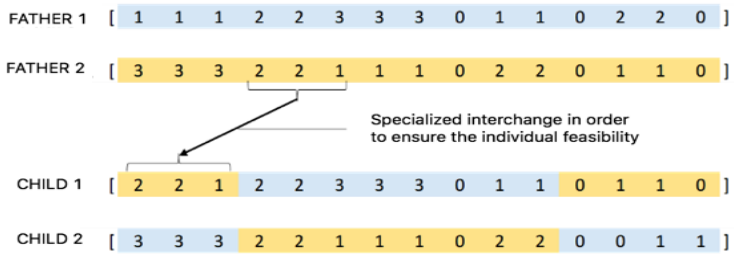


Fig. 2. Crossover operator for product allocation problem.

Table 1. Solution individuals in the Pareto frontier.

Individual	Time	Cost	Individual	Time	Cost
1	40094	22549	7	39893	22600
2	40001	22577	8	40045	22557
3	39846	22729	9	39913	22591
4	40004	22573	10	39972	22578
5	39920	22579	11	39855	22723
6	40009	22563	12	39888	22722

The crossover operator is performed by determining two points randomly selected, from which the parents' genetic information is exchanged to the children, as presented in Figure 2. To ensure the genetic feasibility of each child, a verify and adjust process is performed after the crossover.

The mutation operator is carried in two parts: in the first part, two positions are selected randomly and if the products in the positions are different, these are are exchanged. If the genes are the same, a new random selection is carried out until they are different. The second part is similar to the previous one, but this ensures that one of the positions randomly selected is empty, which gives the algorithm a greater search capability.

The conservation of the best individuals in the upcoming generations is implicit in the NSGAII algorithm, which ensures the elitism, as mentioned previously. The fitness function is evaluated by calculating the cost and time to allocate the products, as expressed in the objective functions equation presented above.

5 Discussion and Analysis of Results

The NSGAII algorithm was applied to solve the allocation problem in a company with 2500 storage positions and 3 products. The algorithm ran with a population of 100 individuals, 200 generations and a mutation percentage of 0.2. The number of optimal individuals in the Pareto frontier produced by the algorithm is 12 and their objective function values are presented in table 1.

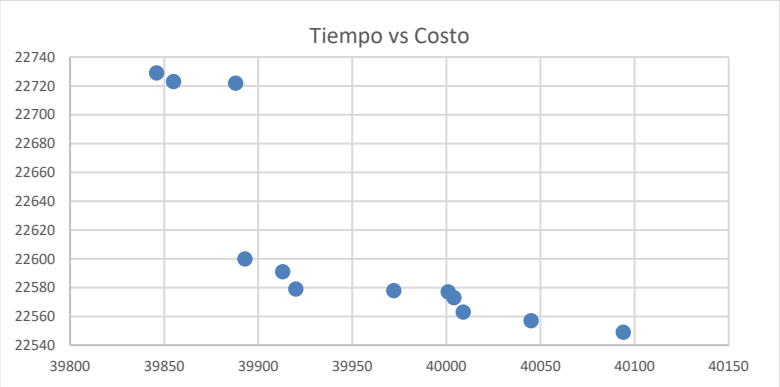


Fig. 2. Pareto Frontier.

Those individual can be graphically presented in order to observe the Pareto frontier behavior in relation to the two objective functions, as presented in Figure 2.

From figure 2 it can be observed that both objective functions are contradictory, that is, the individual that generates the allocation plan with the best costs also produces the worst cost and conversely. In this way, all individuals in the Pareto frontier are optimal combination of the two objectives and the decision maker can select the one he considers that best suits its preferences and also can be sure that no solution is worse or better than any other.

Each of these individuals generates a product allocation plan in the warehouse, indicating what type of product should be storage in each space, as it is defined by the chromosome representation in the genetic algorithm. In this way, it is easy for the decision maker to convert the answer of the algorithm(the individuals) into the warehouse allocation plan. Because to the size of the table that represent each individual solution (a vector with 2500 positions), their presentation is omitted in this paper.

6 Conclusions

This work presented a product allocation model that contemplates the simultaneous optimization of two fundamental objectives for the correct operation and performance in warehouses, which are the costs and the time to fulfil the orders and make the operations. In this way, for the operative plans, managers can find a relationship between these two objective functions that ensures a combination that optimizes both criteria at the same time.

Due to the complexity of the proposed model and the disadvantages of the classic methods to solve multiobjective problems, a genetic algorithm NSGAII was developed, which used an easy-to-understand representation that facilitates the algorithm operation and the conversion of the solutions into the product allocation plan in warehouses. The algorithm ran according to it was expected, generating the Pareto frontier of optimal individuals, from which the decision maker can select any solutions, with the assurance of using an optimal allocation plan in relation to the cost and service time in the warehouse.

As future research work, it is recommended to integrate this model into an information system that not only allows the development of optimal allocation plans, but also provide guidance to workers about the location of the products. In addition, as future research lines, it is suggested to analyze other objective functions to optimize, such as the required area for the allocation plans in the warehouse and other variables, such as costs and times of order preparation.

References

1. Arango-Serna, M.D., Zapata-Cortes, J.A., Pemberthy, J.I.: Reestructuración del layout de la zona de picking en una bodega industrial. Revista Ingeniería Universidad de los Andes, 32, pp. 54–61 (2010)
2. Ballesteros-Riveros, F.A, Arango-Serna, M.D., Adarme-Jaimes, W., Zapata-Cortes, J.A.: Storage allocation optimization model in a Colombian company. DYNA, 86(209), pp. 255–260, (2019)
3. Chopra, S., Meindl, P.: Administración de la cadena de suministro. Pearson Education, Mexico (2013)
4. Correa, C.A., Bolaños, R.A., Molina, A.: Algoritmo multiobjetivo NSGA-II aplicado al problema de la mochila. Scientia et Technica, 14(39), pp. 206–211 (2008)
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.A.: Fast y elitist multiobjective genetic algorithm: NSGA-II. In Proceeding IEEE Transactions on Evolutionary Computation, 6(2), (2002)
6. De Koster, R., Le-Duc, T., Roodbergen, K.J.: Design and control of warehouse order picking: a literature review. Eur. J. Oper. Res., 182, pp. 481–501 (2007)
7. Departamento Nacional de Planeación: Encuesta Nacional Logística 2018 (2018)
8. Fonseca, C.M., Fleming, P.J.: An overview of evolutionary algorithms in multiobjective optimization. Evolutionary Computing, 3(1), pp. 1–16 (1995)
9. Frazelle, E., Soho, R.: World-class warehousing and material handling. McGraw-Hill Education (2016)
10. González, D.L.: Metaheurísticas, optimización multiobjetivo y paralelismo para descubrir motifs en secuencias de AND. Universidad de Extremadura (2013)
11. Hu, W., Wang, Y., Zheng, J.: Research on warehouse allocation problem based on the artificial bee colony inspired particle swarm optimization (ABC-PSO) algorithm. International Symposium on Computational Intelligence and Design, 1, pp. 173–176 (2012)
12. Kovács, A.: Optimizing the storage assignment in a warehouse served by milkrun logistics. International Journal of Production Economics, 133(1), pp. 312–318 (2011)
13. López, A., Zapotecas, S., Coelho, L.C.: An introduction to multiobjective optimization techniques, In Evolutionary Multiobjective Optimization: Theoretical Advances and Applications, pp. 7–32. Springer (2009)
14. Quintanilla, S., Pérez, Á., Ballestín, F., Lino, P.: Heuristic algorithms for a storage location assignment problem in a chaotic warehouse. Engineering (2015)
15. Ramtin, F., Pazour, J.A.: Product allocation problem for an AS/RS with multiple in-the-aisle pick positions. In IIE Transactions, 47(12), pp. 1379–1396 (2015)
16. Sanei, O., Nasiri, V., Marjani, M.R., Moattar, H.S.M.: A heuristic algorithm for the warehouse space assignment problem considering operational constraints: with application in a case study. pp. 258–264 (2011)
17. Shenfield, A., Fleming, P.J., Alkarouri, M.: Computational steering of a multi-objective evolutionary algorithm for engineering design. Engineering Applications of Artificial Intelligence, 20, pp. 1047–1057 (2007)

18. Tompkins, J.A., White, J.A., Bozer, Y.A., Tanchoco, J.M.A.: Facilities planning. Wiley, (2010)
19. Van Wijk, A.C.C., Adan, I.J.B.F., van Houtum, G.J.: Optimal allocation policy for a multi-location inventory system with a quick response warehouse. Operations Research Letters, 41(3), pp. 305–310 (2013)
20. Zapata-Cortes, J.A., Arango-Serna, M.D., Serna-Urán, C.A., Adarme-Jaimes, W.: Mathematical model for product allocation in warehouses. In: García-Alcaraz J., Sánchez-Ramírez C., Avelar-Sosa L., Alor-Hernández G. (eds) Techniques, tools and methodologies applied to global supply chain ecosystems. Intelligent Systems Reference Library, 166, Springer, Cham (2020)
21. Zapata-Cortes, J.A., Arango-Serna, M.D., Adarme-Jaimes, W.: Herramientas tecnológicas al servicio de la gestión empresarial. Avances en Sistemas de Información, 7(3), pp. 87–102 (2010)
22. Zapata-Cortes, J.A.: Optimización de la distribución de mercancías utilizando un modelo genético multiobjetivo de inventario colaborativo de m proveedores con clientes. Universidad Nacional de Colombia, (2016)

ISSN 1870-4069

Convolutional Neural Network in a Pseudo-Distributed Environment for Classification of Chest X-Ray Images of Patients with Pneumonia

Alexandra K. Medrano Roldán, Julia P. Sánchez Solís, Vicente García Jiménez, Rogelio Florencia Juárez, Gilberto Rivera Zárate

Universidad Autónoma de Ciudad Juárez,
Mexico

al137706@alumnos.uacj.mx,{julia.sanchez, vicente.jimenez, rogelio.florencia, gilberto.rivera}@uacj.mx

Abstract. In recent years, there has been an increase in the volume of medical data, generating hundreds of terabytes (TB)/petabytes (PB) of data from different sources. This has led to the emergence of innovative technologies such as Apache Spark, which is a framework that allows the analysis of data in memory based on distributed processing. However, since it is a relatively new technology, both Spark and the other tools that have been developed as a complement, do not have orderly and updated documentation. In this project, a convolutional neural network was implemented in a pseudo-distributed environment for the automatic classification of chest X-Ray images of patients with pneumonia using the Dist-Keras library. Thus, it was possible to explore how the convolutional neural network behaves in Spark as the size of the database increases. While the time was showing an increase as the database grew, the accuracy, precision and sensitivity metrics showed a non-stable behavior.

Keywords: Spark, convolutional neural networks, Dist-Keras.

1 Introduction

This work addresses the behavior of a Convolutional Neural Network (CNN) in a pseudo-distributed environment as the size of the database analyzed increases. The pseudo-distributed environment was configured with Spark, while the database contains chest X-Ray images of patients with and without pneumonia.

In order to analyze the behavior of the CNN in the pseudo-distributed environment, ten data sets were created, from the original database, of different sizes starting with 10% and increasing from ten in ten to 100%. Likewise, to avoid the imbalance of the classes during the analysis, the number of images was reduced from 5,856 to 3,166, where in each data set, there were half of the images belonging to the class with pneumonia and the other half without pneumonia.

Afterwards, the images were normalized and transformed to be displayed in a csv file. The CNN was trained and evaluated with each data set for analysis, so the results correspond to the average of ten executions. Because of the necessary equipment to simulate a distributed environment was not available, the CNN was implemented in a single computer, that is, in a pseudo-distributed environment.

This paper is organized as follows. Section 2 presents a brief summary of some related works. Section 3 shows a description of CNNs. Section 4 presents the evaluation metrics used in this work. Sections 5 and 6 give a short description of Apache Spark and Dist-Keras, respectively. Section 7 shows the implemented method. Section 8 presents the results and discussions. Finally, Section 9 concludes this paper.

2 State of the Art

This section describes some works that deal with the analysis of data, particularly images in a distributed environment.

In [3], a set of medical applications within a Grid Infrastructure is described as well as the Medical Applications on a Grid Infrastructure Connection (MAGIC-5) project. In this work, the authors propose to improve the performance of disease detection programs, such as those in MAGIC-5, by allowing remote access to hundreds of data generated in different hospitals. The MAGIC-5 tool contains algorithms that allow the analysis of: mammograms, for the detection of breast cancer; Computed Tomography (CT) Scans, for the detection of lung cancer; and positron emission tomography, for the early diagnosis of Alzheimer.

In [14], an architecture, which encompasses the parallel image processing in the cloud, as well as a mass data processing engine in Hadoop, are presented. The algorithms that were used for the experiments were: Discrete Fourier Transform (DFT), face detection and template comparison. The results of the experiments showed that the design of this architecture can handle a huge number of large images and videos. However, there were problems regarding the distribution of information and response time.

In [15], a toolkit, called Kira, was built for the processing of astronomical images using Apache Spark and running on Amazon Elastic Compute Cloud (Amazon EC2). The authors reported that the experience with Kira showed that applications in the Apache Spark area are an alternative for multitasking scientific applications and that, specifically, Apache Spark can be easily integrated with existing libraries.

There are three main differences between the works mentioned above and the present project. First, in this work the analysis was applied to a database with images of patients with and without pneumonia. Second, the image classification method selected was a CNN. Finally, the third difference is that the algorithm was implemented in a pseudo-distributed environment, using Apache Spark.

3 Convolutional Neural Network

A CNN is a deep and feed-forward artificial neural network, where a hierarchical structure is maintained through the learning of representation of the internal characteristics. In addition to generalizing them in image problems, CNNs have also achieved results in problems related to natural language processing and speech recognition [10].

The main idea of CNNs is to devise a solution to reduce the number of parameters allowing a network to be deeper with fewer parameters [2]. The three most common types of layers in a CNN are: convolution layer, pooling and Rectified Linear Units (ReLU), as well as the fully connected layer [1].

In this project, the data was divided into two data sets, one for training and the other for testing. The CNN learned from with the training data set, and then its performance was evaluated with the testing data set.

4 Evaluation Metrics

The evaluation of the model is the most important step in the development of any machine learning solution since it determines if it is necessary to review the previous steps before continuing [12]. The metrics used in this project are described below:

- **Accuracy:** The accuracy of a classifier is the percentage of a set of test examples that were correctly classified [4]. It is defined as:

$$Accuracy = \frac{TP + TN}{P + N},$$

  where TP is the number of positive cases that are truly positive, and TN is the number of cases that the test marks as negative and they are truly negative. Meanwhile, P and N are the total positive and negative examples, respectively.

- **Sensitivity:** The sensitivity or true positive rate (TPR) is the probability of positive examples correctly classified [4]. It is defined as:

$$Sensitivity = \frac{TP}{TP + FN},$$

  where FN is the number of positive cases that were classified as negative.

- **Specificity:** The specificity or rate of true negatives (TNR) is the probability of correctly classified negative examples [4]. It is defined as:

$$Specificity = \frac{TN}{TN + FP},$$

where FP is the number of negative cases that the algorithm classified as positive.

- **Precision:** The precision is the result of dividing the true positive values (TP), among all the positive classifications (P') [8]. The formula is as follows:

$$Precision = \frac{TP}{TP + FP}.$$

5 Apache Spark

Apache Spark is a data analysis system in memory based on distributed processing. It is built on Hadoop and, therefore, uses Hadoop Distributed File System (HDFS) as a file system for data storage. However, there are differences between Hadoop and Spark. One of them is that Hadoop stores the data on disk to run the analysis, while Spark uses a cache mechanism in memory to store the data and process it [13].

6 Dist-Keras

Distributed Keras [7] is a distributed deep learning framework. It was developed based on Apache Spark and Keras [5], with the aim of significantly reduce training with distributed machine learning algorithms.

The distributed machine learning approach that follows is the parallel data paradigm, in which copies of the model are held in different trainer distributed on different nodes, although they may be on the same machine. In addition to that, the data set is divided in such a way that each replicated model can be trained in a different partition of the complete data set [6].

7 Method

In this section, the method used to carry out the project is presented.

7.1 Experimental Approach

In this project, the database with chest X-Ray images that were collected and labeled in [9] was used. This database has 5,856 images of chest X-Ray of children, where 4,273 were from patients with pneumonia, while 1,583 were from patients with a healthy diagnosis. Of the chest X-Ray images with pneumonia, 2,780 were due to bacteria and 1,493 were due to virus.

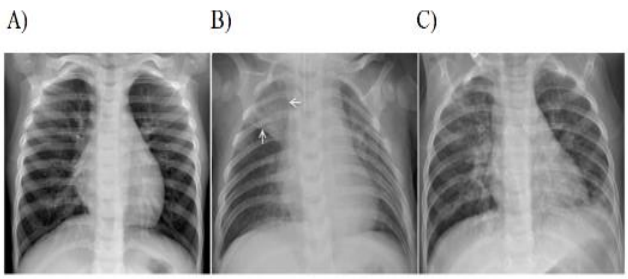


Fig. 1. A) Normal, B) Bacterial pneumonia, C) Viral pneumonia [9].

The chest X-Ray images were taken as part of the usual clinical care of the patients and reviewed to eliminate low quality or illegible scans, then they were qualified by experts. In Figure 1 there are three images of different patients belonging to the three categories of pneumonia: without, bacterial and viral.

7.2 Data Preparation

For the preparation of the data, first, because the images had different resolutions, it was decided to scale all the images to 100x100.

On the other hand, of the 5,856 images, 238 were in color, that is, they had the three RGB channels, while the rest had only one-color channel, in other words, the rest of the images were gray-scale. To normalize the images, those 238 that were in color were changed to gray-scale.

Then, in order to adapt the database to the image classification algorithm, each image was represented as a single vector within a csv file. In this file each row represents a different image. The first column of each row is the class to which the image belongs, while the rest of the columns correspond to the pixels. In the case of images with normal diagnosis, the label was 0, while, for the chest X-Ray images of patients with pneumonia, it was assigned the label 1. Because the final set of images had a resolution of 100x100 pixels, each image was transformed to a vector with 10,001 columns: one for the label and the remaining 10,000 for each of the pixels in the image.

For the classification, it was decided to consider only two classes: with normal and pneumonia diagnosis, grouping patients with pneumonia due to bacteria and viruses into a single class: pneumonia.

For the final preparation of the data set, the images in the folder val were integrated into the ones of test and train. Later, in order to avoid the imbalance of classes for each data set, the number of images per folder was reduced. The reduction was done randomly, where the number of images that would make up 100%, was defined according to the number of images in the class with the lowest presence in the database: without pneumonia or normal. Once the 100% data sets were defined, the other data set were obtained by percentages, as it is shown in the Table 1. In order to be able to use the images with the classifier, the code fragment for image preprocessing found in [6]

Table 1. Database division.

Percentage	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Test	48	96	144	192	241	289	337	385	433	484
Train	268	536	804	1072	1340	1608	1876	2144	2412	2682

Table 2. Confusion matrix for the classes with and without pneumonia.

		Predicted class		
		With pneumonia	Without pneumonia	Total
Actual Class	With pneumonia	TP	FN	P
	Without pneumonia	FP	TN	N
	Total	P’	N’	P + N

was used. In this code section, pixel values were normalized. Instead of having values between 0 and 255, they were normalized to range between 0 and 1.

7.3 Modeling

The image classification method selected for this work was a CNN, implemented in a pseudo-distributed environment using Apache Spark 2.4 together with Hadoop 2.7. The computer used has as operating system Ubuntu 18.04.2 LTS, has 12 GB of RAM and an Intel Core i7-2600K processor, with 3.40 GHz x 8. It was necessary to create a virtual environment in Anaconda installing Python 3.5.5.

The CNN was implemented based on the image processing example of the Modified National Institute of Standards and Technology (MNIST) database in [6], and in order to implement this network in a pseudo-distributed environment, the Dist-Keras library was used. Specifically, the distributed optimizer Asynchronous Elastic Averaging Stochastic Gradient Descent (AEASGD).

7.4 Evaluation

The metrics evaluated were time, accuracy, precision, sensitivity and specificity. Because the Dist-Keras library can only evaluate accuracy and time, the missing metrics: precision, sensitivity, and specificity were implemented in the library code. For the evaluation of the metrics the class with pneumonia was defined as positive, while the negative class was that of patients without pneumonia. With the above, the confusion matrix would be as shown in Table 2.

8 Results and Discussions

In this section, are shown the values of the metrics as well as their analysis. The metrics includes: time, accuracy, sensitivity and specificity, precision and, finally, the ROC Curve.

8.1 Time

When evaluating the training time of the CNN in the pseudo-distributed system, the overall results obtained from the ten executions per data set were averaged and are shown in Table 4, with their respective standard deviation.

When analyzing the results, in a matter of time, the CNN in Spark behaves as expected: the greater the size of data, the greater the training time. Table 3 shows the increase in time as the size of the database increases. The times obtained as the number

Table 3. Increase in time.

Percentage	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Test	53.1	85.8	112.2	170.1	184.3	243.0	340.6	345.0	393.9	374.5
Train	-	536	804	1072	1340	1608	1876	2144	2412	2682

of images was increased, can be modeled with the function 1:

$$t(n) = 0.13n + 4.11\,, \tag{1}$$

where t is the time in seconds and n the number of images. The function was obtained through the method of least squares. The function indicates that time grows linearly as the number of images increases, meaning that a larger number of images would not imply large or exponential growth.

8.2 Accuracy

When evaluating the accuracy of the CNN after the training, the overall results obtained from the ten executions per data set were averaged and are shown in Table 4, with their respective standard deviation.

In the case of accuracy, not much variability is shown, in relation to the other metrics as the size of the database increases. The range is maintained between 0.45 and 0.55, or what would be between an interval of 45 to 55% accuracy on average. Indicating that, for the most part, it only classifies well 50% of the cases of the total. It is important to mention here that the highest values were obtained with the 10% data set, that is, the data set with the smallest size.

8.3 Sensitivity and Specificity

The average of values, as well as their standard deviation, obtained when evaluating the Sensitivity and Specificity of the CNN is shown in the Table 4.

For the sensitivity, values were obtained between 0.27 and 0.65, or what would be between an interval of 27 to 65% of sensitivity. There is not such a clear trend depending on the size of the database as there was with the metric accuracy. Specificity, a metric that has the same relationship as sensitivity, but with respect to the negative class, or in this case, with respect to the class to which patients without pneumonia belong, obtained values ranged from 0.31 to 0.67, or from 31 to 67%.

Table 4. Results.

Percentage	Time		Accuracy		Sensitivity		Specificity		Precision	
	M	SD	M	SD	M	SD	M	SD	M	SD
10%	53.05	0.774	0.550	0.151	0.554	0.370	0.546	0.450	0.709	0.266
20%	85.83	8.653	0.466	0.062	0.460	0.393	0.471	0.374	0.422	0.117
30%	112.2	1.904	0.510	0.090	0.400	0.391	0.619	0.437	0.627	0.239
40%	170.1	58.79	0.449	0.053	0.266	0.333	0.632	0.369	0.337	0.259
50%	184.3	15.67	0.477	0.052	0.451	0.443	0.503	0.436	0.470	0.249
60%	243.0	67.42	0.476	0.056	0.641	0.409	0.310	0.388	0.522	0.184
70%	340.6	59.71	0.516	0.068	0.647	0.441	0.383	0.422	0.429	0.188
80%	345.0	54.96	0.520	0.048	0.640	0.445	0.399	0.420	0.474	0.273
90%	393.9	53.44	0.505	0.048	0.455	0.419	0.555	0.404	0.522	0.086
100%	374.5	20.61	0.498	0.014	0.322	0.468	0.574	0.464	0.396	0.306

Considering both sensitivity and specificity, we can see that, in general, the CNN was able to better classify images of patients without pneumonia than of patients with pneumonia. This could be because, among the images with pneumonia, there were those of patients infected by viruses and bacteria, which were grouped into a single class: pneumonia.

8.4 Precision

The mean values, as well as the standard deviation, are shown in Table 4 and were obtained when evaluating the precision of the ten executions of the CNN.

In the case of the precision metric, the values ranged from 0.34 to 0.71, or what would be between 34 and 71%. The average values were higher, although these will have to be taken with reservations, this due to the behavior of the metric to produce high values in situations such as, for example, suppose you have 30 examples of patients with pneumonia and only one of them is classified correctly and with no *PF* (Positive false) values, then the precision value would be 1 or 100%, this despite there are 29 images of patients with pneumonia that were classified incorrectly.

8.5 ROC Curve

In the ROC curve, the sensitivity values are on the y-axis, while the values of 1-specificity are on the x-axis. The graph is shown in Figure 2.

In Figure 2, it can be seen that there are some data sets in the desirable area, such as the 10%, 30%, 70%, 80% and 90% data sets. On the other hand, the rest of the data sets tended to classify the images in a more random way.

The random behavior in the metrics could be since only one epoch was defined in the training for the ten data sets, for memory reasons, despite the increase in the size of

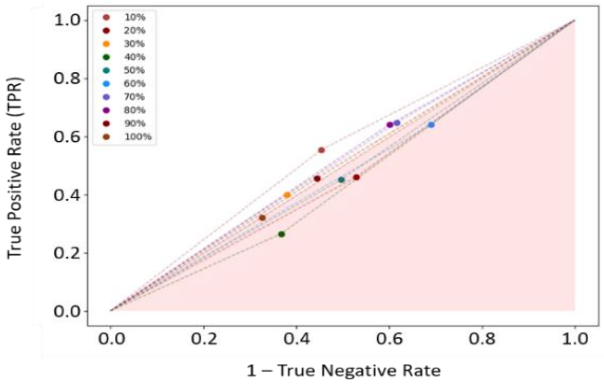


Fig. 2. ROC Curve.

the training data sets. This behavior of the algorithm is logical, since the larger the database, the more complex it becomes. Therefore, it is necessary to increase the number of epochs as the size of the data set increases. In addition, there is the possibility of the noisy factor that can be present in the database from the moment the images were captured. Regarding the noisy factor, although it is a latent problem, it was left like that in order to focus on the behavior of the CNN in Spark.

9 Conclusions

In this project, a Convolutional Neural Network (CNN) was implemented in a pseudo-distributed environment in Spark using the Dist-Keras library. The CNN was used for the automatic classification of images from the database of patients with and without pneumonia taken of [11]. The configuration of the CNN was obtained from the example applied to the MNIST database [6].

Database was divided in ten image sets, each one with different size in order to explore the behavior of the CNN as the size changes. The metrics of precision, accuracy, sensitivity and specificity were used to evaluate the CNN. The CNN was run ten times on each image set to perform classification tests.

Although the database increased, the CNN was able to continue the training and classification process, with an increase in time as expected due to the increase in the size of the database. On the other hand, the metrics values became questionable, due to the great variability of them as the size of the database changes.

After the relevant tests, it was confirmed that the training time of the CNN increases as the size of the database is extended. For future work, it is recommended to increase the number of times for the CNN's training, since as the size of the database increases, the analysis process becomes more complex. Similarly, it is recommended to implement the CNN in a cluster, to see what the performance of the CNN in Spark is as the number of slave nodes increases. Finally, images of patients with bacterial and

viral pneumonia could be separated into two distinct classes, with the intention of decreasing the number of false negatives.

Acknowledgment. The authors would like to acknowledge the financial supports from the Mexican PRODEP (Project UACJ-PTC-423).

References

1. Aggarwal-Charu, C.: Neural networks and deep learning: A textbook. Yorktown Heigths, Springer (2018)
2. Aghdam-Hamed, H., Jahani-Heravi, E.: Guide to convolutional neural networks: A practical application to traffic-sign detection and classification. Springer (2017)
3. Belloti, R., Tangaro, S., Cerello, P., Bevilacqua, V.: Distributed medical images analysis on a grid infrastructure. Future Generation Computer Systems, 23, pp. 475−484 (2007)
4. Cali, C., Longobardi, M.: Some mathematical properties of the ROC curve and their applications. Ricerche di Matematica, Springer Nature, 13, pp. 391−402, (2015)
5. Hermans-Joeri, R.: CERN IT-DB Distributed Keras: Distributed deep learning with apache spark and keras. Github Repository (2016)
6. Hermans-Joeri, R.: Distributed Keras: Introduction. https://joerihermans.com/work/distributed-keras/ (2019)
7. Juba, B., Le-Hai, S.: Precision-recall versus accuracy and the role of large data sets. https://pdfs.semanticscholar.org/7abb/63a5cb77a0ada993cfe2328f38689431e2da.pdf. (2019)
8. Keras Documentation: Keras: the Python deep learning library (2019)
9. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172, pp. 1122−1131 (2018)
10. Kumar- Manaswi, N.: Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition with Tensorflow and Keras. Apress (2018)
11. Mooney, P.: X-Ray Images (Pneumonia). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/home (2018)
12. Ramasubramanian, K., Singh, A.: Machine learning model evaluation. Machine Learning Using R, Apress (2016)
13. Srinivasa, K.G., Siddesh, G.M., Srinidhi, H.: Apache spark computer communications and networks. Springer International Publishing (2018)
14. Yan Yuzhong, Huang Lei: Large-scale image processing research cloud. In: The Fifth International Conference on Cloud Computing, GRIDs, and Virtualization, pp. 88−93 (2014)
15. Zhang, Z., Barbary, K., Nothaft, F.A., Sparks, E., Zahn, O., Franklin, M.J., Patterson, D.A., Perlmutter, S.: Scientific computing meets big data technology: An astronomy use case. In: IEEE International Conference on Big Data, University of California Press (2015)

Thematic Section
“Machine Learning for Healthcare: Modeling, Analysis and Computer Simulation”

Alfonso Rojas Domínguez
Matías Alvarado (eds.)

ISSN 1870-4069

Editorial for Thematic Section "Machine Learning for Health Care: Modeling, Analysis and Computer Simulation"

Effective computer applications targeting healthcare problems can be traced back to at least the second half of the 20th century. During the first half of the 21st century, this area of application has experienced a revolution, driven by the advent and development of advanced Machine Learning techniques, and supported by the ever more powerful computer hardware, electronics, and the increasingly prevalent global-communications infrastructure. All of these are bringing the medical fields, and computer science, closer together than they have ever been.

This thematic section collects research derived from the current intense interaction between computer scientists working on artificial intelligence, machine learning and big data, and clinicians / medical researchers working on the different branches of health care. The field of MLHC supports the advancement of data analytics, knowledge discovery, and the meaningful use of complex medical data by fostering collaboration and the exchange of ideas between these communities.

The world health challenges on cancer, malaria and retinal diseases growth, that involve intense data analysis as well as precise but flexible modeling, concern this volume of AI solutions.

In "CAD of Breast Cancer: a decade-long review of techniques for Mammography Analysis", Rojas et al. discuss the paradigm shift observed within the machine learning research community (from feature-engineering to feature-learning) and how it is reflected on the design of computer-aided systems for mammography analysis used against breast cancer.

In "Machine Learning Techniques for Diagnosis of Breast Cancer", A. Rojas offers an overview of the recent machine learning techniques for diagnosis of breast cancer, presented from a perspective of the work that he has carried out through several years. The techniques used in the two primary tasks for the early diagnosis of breast cancer (mass detection and mass classification) are discussed.

In "Evaluation of breast cancer by infrared thermography", Morales-Cervantes et al., discuss the analysis of thermograms of patients with suspected breast cancer. The asymmetry of a thermal score (combination of the amount of vascularization and surface temperature) between the breasts of a patient is indicative of anomalies. The automated method presented achieves higher sensitivity (100%) and specificity (68.68%) than an expert oncologist, on 206 test thermograms.

In "Cancer metastasis and the immune system response: CM-IS modeling by Ising model", Alvarado and Arroyo introduce this highly complex biological process, of top interest for cancer diagnosis and therapy. The strength of the immune system response against cancer correlates with the success of the cancer growth. Their interaction is formalized by the Ising model (a classic model for emergent-interaction phenomena) and simulated by means of an agent-based environment.

In "Automatic Cropping of Retinal Fundus Photographs using Convolutional Neural Networks", González-Briceño et al. present a segmentation method based on Deep

Convolutional Networks for cropping of the Region of Interest in Retinal Fundus Photographs (used by physicians to diagnose ocular diseases). The proposed method achieves high accuracy levels (up to 98%) on test images.

In "Random forest and deep learning performance on the Malaria DREAM sub challenge one", D. Barradas-Bautista introduces his machine learning solutions to build a predictor of the clearance and concentration of Artemisin (the standard antimalarial drug) based on multivariate data with over 5000 features. A random forest model allows the ranking of features, which could become new drug targets, and a fully connected Deep Network model achieves a prediction accuracy of 72.20%.

Alfonso Rojas Domínguez (CONACYT Research Fellow, Tecnológico Nacional de México – Campus León, Mexico)
Matías Alvarado (Centro de Investigación y de Estudios Avanzados del IPN, Mexico)
Guest Editors

June 2020

ISSN 1870-4069

CAD of Breast Cancer: A Decade-Long Review of Techniques for Mammography Analysis

Alfonso Rojas Domínguez, Héctor Puga,
Manuel Ornelas Rodríguez, Itzel Guerrero Gasca

Tecnológico Nacional de México,
Campus León,
Mexico

{alfonso.rojas, lezti10.ig}@gmail.com

Abstract. Breast cancer is the most common type of cancer among women. Early detection is essential to reduce mortality by breast cancer, and mammography is the main tool for detection and first diagnosis of breast cancer. However, mammogram interpretation is a difficult task, and up to 30% of visible cancers in mammograms are missed by human readers. For this reason, Computer-Aided Detection/Diagnosis (CAD) systems have been developed since the 1990's to ameliorate this problem. On the other hand, the popularity of some Machine Learning (ML)/Artificial Intelligence techniques has increased dramatically in the last decade due to technology development leading to improved performance. Naturally, these techniques have also been introduced into CAD systems for mammography analysis. In this work, a review of the techniques employed in CAD systems is presented; this review is focused on the paradigm shift observed within the ML research community and how this has been reflected in the design of CAD systems for breast cancer based on mammography analysis. Through this work, the reader may gain valuable insights into this field.

Keywords: deep learning, mammography analysis, breast cancer, CAD.

1 Introduction

Breast cancer is the most common[1] cancer in women worldwide with over 2 million new cases in 2018, contributing 25.4% of the total number of new cases diagnosed in that year.[2] Early detection of breast cancer can help to increase the 5-year survival rate. The primary technology for breast cancer screening is mammography. Double-reading from independent expert radiologists is recommended to minimize detection misses; however, this practice represents additional costs and workload. Although there are ethical and technical difficulties that prevent Computer-Aided Diagnosis (CADx) systems from completely replacing human

[1] Excluding non-melanoma skin cancer, which is extremely common but often curable.
[2] https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data

experts, these systems can provide second-, or third-opinions to support the decisions of radiologists based on mammography [8].

In this work we present a review of the most pertinent research about Machine Learning (ML, a subset of Artificial Intelligence) based CADx of breast cancer that has been published in the past ten years. Our review is limited to the techniques that have been developed for the analysis of mammograms and is presented from the perspective of ML practitioners, rather than from that of a radiologist or a health care expert. A search of the literature in this topic produces thousands of results. Necessarily, only the most cited papers and from the publications with the highest impact factors were included in this review. Also, only studies that experiment on mammography, either as digital/digitized mammograms or as Digital Breast Thomosynthesis (DBT), were considered. A number of recent surveys exist on the topic of ML/AI in medical imaging [22,31,23], which were also examined for contributions relevant to the topic of mammography analysis for CADx of breast cancer.

Our review is organized chronologically into two periods: the *Feature Engineering* period (prior to 2015) and the *Feature Learning* period (2015 to date). Also topically, according to the main tasks tackled by CAD systems: detection of abnormalities (masses or microcalcifications), and classification of abnormalities. Thus, our work is presented in two main sections, with subsections corresponding to the mentioned tasks.

2 Background

Before Deep Learning's (DL) near hegemony in almost every area of practical application of ML [27], the research on medical image analysis in general, and on mammography analysis in particular, was organized by tasks rather than by techniques. Thus, different techniques from ML / image processing were used for the detection of breast masses, while other (sometimes similar but often quite different) techniques were developed for their classification into benign/malignant classes, etc. The advent of DL with the AlexNet of Krizhevsky et al. in 2012 [26,2], and its further impressive results in many of the areas of application of ML, circa 2015, brought with it a radically different paradigm in the way that medical image analysis can be addressed. Thus, in most surveys published after 2016, two clearly delineated approaches for the use of ML in CAD are present.

The two paradigms are illustrated in Fig. 1. In the *Feature Engineering* approach, a human expert performs feature design and selection, based on problem data from which specific problem knowledge is extracted, and on his/her own engineering expertise. Afterwards, to be classified, a data instance needs to be preprocessed to extract the engineered features, and these can be fed to a simple classifier[3] such as an Artificial Neural Network (ANN) a Support Vector Machine (SVM) [12], etc.

[3] Said classifier must be trained beforehand on a dataset of extracted features; this process has been disregarded for the sake of simplicity in our exposition.

In the *Feature Learning* paradigm, the problem data is used to train a Deep Neural Network that includes a Feature Extraction Stage and a so-called Classification Head (a classifier embedded within the deep network). Afterwards, a data instance can be input directly into the trained network, to be classified. Notice that substantially more data is required to perform automated feature learning than what is usually required for feature engineering[15]. The reason for this is that humans inherently possess generalization capabilities that allow us to discover patterns from, and to transfer prior knowledge into, new problems; on the other hand, a network must learn everything from scratch, which can only do by examining large amounts of training data [27,7].

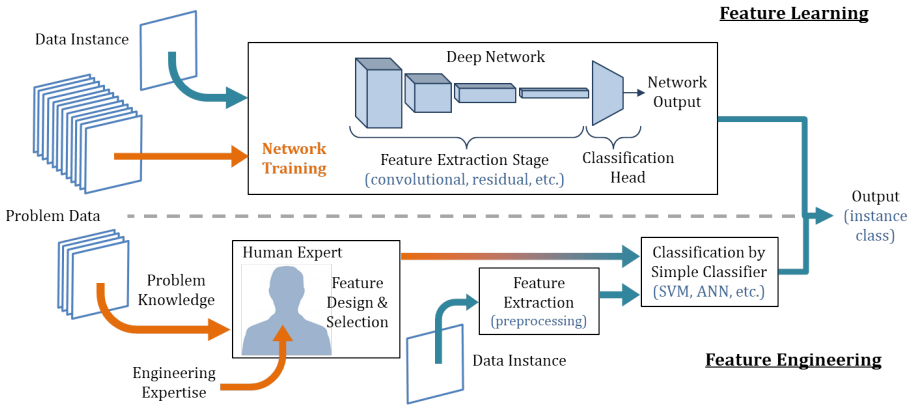


Fig. 1. Two paradigms of ML-CAD: in Feature Engineering the orange arrows represent the data flow for feature design & selection, while in aqua is the classification of unseen instances. In Feature Learning, a Deep Network is trained on large amounts of data (orange), afterwards the network will perform feature extraction & classification of new instances (aqua).

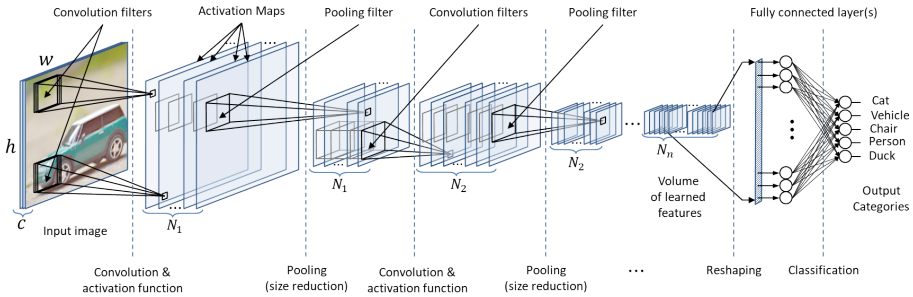


Fig. 2. Convolutional Neural Network for classification.

With few exceptions, the features that are manually engineered by human experts cannot be applied to different problems. In other words, features engineered for breast-mass detection will not be useful for mass segmentation, nor for mass characterization, etc.

On the other hand, the features ultimately learned by a Deep Network, or high-level features, are based on other lower-level features, hierarchically

generated by different layers of the network. The features in the first layers describe abstract shapes (lines, edges, gradients, etc.); the features in the last layers correspond to the specific structures of interest, such as spiculated masses or fuzzy boundaries indicative of malignancy. Thus, low-level features possess a degree of generality, so that to some extent the knowledge acquired by the network can be *transferred* to other networks designed for different problems. This is known as Transfer Learning (TL) [6], and is an important advantage of DL over other ML techniques.

Regarding the type of networks that can be employed at the core of the Feature Learning paradigm, we can list a variety of architectures [7]: Convolutional Neural Networks (CNNs), Residual Networks, Recurrent Networks, Deep Autoencoders [21], Deep Belief Networks [20], etc. The most commonly employed type of network for mammography analysis is the CNN [1] (Fig. 2). A CNN is characterized by being formed by several *convolutional* layers, that contain a number of trainable convolution filters that produce so-called activation maps.

3 Feature Engineering Period (2009-2015)

Detection of breast masses by means of 2D mammography is hindered by dense parenchyma, which can also generate false positive (FP) detection because the overlapping tissue may mimic lesions. One alternative is the use Digital Breast Tomosynthesis, patented in 1999[4], although initial experiences and comparisons between conventional mammography and DBT were being performed circa 2007 [19,35]. DBT produces a set of 2D images or 'slices' from which a 3D voulume of the breast can be reconstructed.

Overlapping of the parenchyma is reduced and this may offer higher sensitivity than regular mammograms (and reduce recall rate [17]), but with increase in workload. In the Feature Engineering period, several studies describe the developing of early CADe systems to automate the analysis of DBT data [11,9].

Detection of Masses.– In 2008, Chan et al. presented a comparison of three approaches for CADe of masses in DBT [10]. The comparison was carried out on a dataset of 100 DBT cases. The approaches differ in the information that their system employs (i.e. the 3D DBT reconstruction; the backprojected 2D images that form a DBT case; or the combination of those). Detection was based on three-dimensional gradient field analysis and several features that describe a candidate mass. Features included morphological features (volume (3D), area (2D), perimeter, diameter and compactness); statistics of the gray level; run length texture features[14] obtained via the Rubber Band Straghtening Transform [39] and from spatial gray level dependence matrices [18], and a Hessian feature in 2D. The authors reported that the combined approach was superior to the other approaches compared. In a previous work (Rojas-Domínguez & Nandi, 2009) we explored the techniques for segmentation of masses, including the standard features and other proposed features [38].

[4] https://www.massgeneral.org/imaging/services/digital-breast-tomosynthesis.aspx

Detection of Microcalcifications.– Typical techniques employed for detection of microcalcifications vary from simple thresholding to sophisticated methods, such as: the Wavelet Transform and similar techniques, like the Contourlet Transform, that also enable the generation of mammograms in super-resolution [33,47], as well as image denoising and enhancement [16]; the Laplacian of Gaussian (or Difference of Gaussians) filter, used as a *blob* detector; morphological operators; and other ad-hoc methods such as fuzzy-logic, Gabor filters, Zernike moments, etc. A previous work (Rojas-Domínguez & Nandi, 2008) provides a review of the most important of those methods [36]. The review by Elter& Horsch [13] includes a table comparing these techniques in terms of their strengths and weaknesses, while Pak et al. provide a performance-based comparison of these techniques (up to 2015) [33].

Elter & Horsch stated that "half of all missed cancers seem to be missed due to missclasification rather than due to oversight"[13]. This means that CADx systems to characterize lesions are as important as CADe systems that detect those lesions.

Classification of Abnormalities.– The most crucial features for classification of breast masses are their shape and appearance of their margins. In simple terms, benign masses possess clearly defined, smooth margins and round shapes, while malignant masses appear as irregular shapes with diffuse/fuzzy or *spiculated* margins (a spicule is a needle-like structure). Irregular borders indicate that the mass has invaded surrounding tissue, a sign of a metastatic process. Typical results achieve a sensitivity of 86% with 3 FP per image [41]. See our previous work (Rojas-Domínguez & Nandi, 2009) for a discussion of this topic and our proposed robust features for characterization of breast masses [37].

4 Feature Learning Period (2015-to date)

A comprehensive technical review of DL in medical image analysis is provided by Shen et al. that includes details about the most common DL models and an overview of the different applications [42]. A more general overview of DL in medical imaging is [45]. One of the first works that tested the use of DL techniques for mammography analysis is [4]. In 2016 Arevalo et al. studied the automatic classification of breast lesions on a proposed biopsy-proven benchmarking dataset from 344 cases with a total of 736 film[5] mammography views, and manually segmented lesions (426 benign and 310 malignant) [5]. In contrast to previous works where DL models were trained in an unsupervised way [34,24], the authors trained a CNN for feature learning in a supervised fashion, with the peculiarity that instead of performing the classification by means of embedded fully-connected layers, they used an independently trained SVM (nowadays this

[5] That project has been updated to digital mammograms: https://bcdr.ceta-ciemat.es/

practice is far less common). They also used a preprocessing stage with local and global image normalization and data augmentation.[6]

The authors report that their method exhibits improved performance (from 0.787 to 0.822 AUC[7]) when compared to state-of-the-art image descriptors, such as Histogram of Oriented Gradients and Histogram of the Gradient Divergence [32]. Their model also outperformed a set of 17 hand-crafted features that take advantage of additional information from segmentation by the radiologist (notice that these are generic image descriptors):

- Intensity: mean, median, maximum, minimum, standard deviation, skewness, kurtosis.
- Shape: area, perimeter, circularity, elongation, mass centroid, form.
- Texture: contrast, correlation, entropy.

In recent studies, Giger [15] states that both handcrafted and deep-learned features are important, underlining that systems combining both types of features perform the best [3].

Detection of Abnormalities.– In 2016, Samala et al. employed TL on a CNN-based CAD system to improve detection of breast masses in DBT [40]. They pre-trained the more generic layers (first 3 of 4), on a large training set of mammograms and then trained the more specific layer on DBT data. The CNN-based CAD system was compared against feature-based CAD including morphological (volume, area, perimeter, longest diameter and compactness), grey level (statistics, contrast and histogram features) and texture features (run-length statistics on the rubber-band [39] of the objects margin). Based on said features, the authors applied linear discriminant analysis-based classifiers to perform FP reduction. They obtained statistically significant improvements of their CNN-based CAD system over their feature-based system. More importantly, they showed that TL can preserve the low level similarities and capture the high-level differences between representations of the masses.

In 2017 Kooi et al. compared a CNN-based system for detection of mammographic lesions against a state-of-the-art, feature-based system relying on a set of 74 manually engineered features [25]. Among these features, those related to Location, Context and Patient features were also provided to the CNN-based system because they were considered complementary to the information that the CNN could extract from the training mammograms. A total of 40,090 mammograms were used for training, and 18,182 for evaluation. The CNN was a scaled-down VGG [43] with 5 convolutional layers plus 2 fully connected layers. The comparison showed no statistically significant differences ($p = 0.2$) using the AUC. However, at high specificity, statistically significant differences were found between the CNN+manual features against their reference state-of-the-art system. Finally, in a comparison against human readers, the performance of the CNN was lower than the average of 3 human readers.

[6] Data augmentation produces variations of the original input data by performing simple geometric transformations such as translations, rotations and mirroring.

[7] Area Under the ROC Curve, a detection performance measure.

Classification of Abnormalities.– The very recent review of Abdelhafiz et al. identifies major challenges in training CNNs for mammography analysis and the most popular solutions that have been employed to overcome them [1]. A major challenge for the training of deep networks is the lack of sufficient labelled data.

Sun et al. presented a graph based scheme of semi-supervised learning (SSL) for CNNs [44]. While CNN training usually requires a large amount of labelled data, their proposed scheme allowed the training of the CNN with only (before data augmentation) 100 labeled ROIs plus 2400 originally unlabeled ROIs that were labeled automatically by a co-training algorithm [17]. Their CNN followed the design of LeCun [29,28] and was formed by 3 convolutional layers. The CNN trained by SSL achieved the best performance (AUC=0.88), followed by an SVM also trained by SSL (AUC=0.85).

Ting et al. [46] recently proposed to employ an interactive detection-based lesion locator by means of an auxiliary network of the type Single Shot Multibox Detector [30], which is a deep network developed for multiple object detection. Their proposal achieves an AUC=0.90. Ting et al. also performed a sophisticated preprocessing, including patch-based data augmentation by rotating and flipping the lesion patches.

5 Conclusion

The effects of the paradigm shift observed in the field of ML extend to those scientific fields in which ML techniques have been successfully applied. The ML-based CAD systems for mammography analysis developed in the last decade, clearly reflect these effects. However, the dominance of DL in this field is not as extensive as in other ML applications. Based on the present review, we can conclude that the *Feature Engineering* (favouring generic image descriptors) and the *Feature Learning* approaches, currently coexist. The best performance is obtained by fusion strategies that combine both. Large high-quality mammography/DBT datasets could promote the maturity of DL techniques in the near future, but these are not easy to compile. In the meantime, the inclusion of complementary information not available in the mammograms, plus such strategies as TL and data augmentation, are essential to the implementation of the most competitive CAD systems.

Acknowledgements. This work was supported by the National Council of Science and Technology of Mexico (CONACYT) under Research Grant: CÁTEDRAS-2598 (A. Rojas) and Postgraduate Scholarship 748219 (Itzel Guerrero).

References

1. Abdelhafiz, D., Yang, C., Ammar, R., Nabavi, S.: Deep convolutional neural networks for mammography: advances, challenges and applications. BMC Bioinformatics 20(11), 281 (2019)

2. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Esesn, B.C., Awwal, A.A.S., Asari, V.K.: The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv:1803.01164 (2018)
3. Antropova, N., Huynh, B., Giger, M.: Deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med. Phys. 44(10), 5162–5171 (2017)
4. Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Lopez, M.A.G.: Convolutional neural networks for mammography mass lesion classification. In: 37th Annual International Conference of the IEEE Eng. in Medicine and Biology Society (EMBC). pp. 797–800 (2015)
5. Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Lopez, M.A.G.: Representation learning for mammography mass lesion classification with convolutional neural networks. Computer Methods and Programs in Biomedicine 127, 248–257 (2016)
6. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning. pp. 17–36 (2012)
7. Bengio, Y., et al.: Learning deep architectures for AI. Foundations and Trends® in Machine Learning 2(1), 1–127 (2009)
8. Birdwell, R.L.: The preponderance of evidence supports computer-aided detection for screening mammography. Radiology 253(1), 9–16 (2009)
9. Chan, H.P., Wei, J., Sahiner, B., Rafferty, E., Wu, T., Roubidoux, M., Moore, R., Kopans, D., Hadjiiski, L., Helvie, M.: Computer-aided detection system for breast masses on digital tomosynthesis mammograms: preliminary experience. Radiology 237(3), 1075–1080 (2005)
10. Chan, H.P., Wei, J., Zhang, Y., Helvie, M.A., Moore, R.H., Sahiner, B., Hadjiiski, L., Kopans, D.B.: Computer-aided detection of masses in digital tomosynthesis mammography: Comparison of three approaches. Medical Physics 35(9), 4087–4095 (2008)
11. Chan, H.P., Wei, J., Zhang, Y., Moore, R., Kopans, D., Hadjiiski, L., Sahiner, B., Roubidoux, M., Helvie, M.: Computer-aided detection of masses in digital tomosynthesis mammography: combination of 3d and 2d detection information. In: Medical Imaging 2007: Computer-Aided Diagnosis. vol. 6514. International Society for Optics and Photonics (2007)
12. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
13. Elter, M., Horsch, A.: Cadx of mammographic masses and clustered microcalcifications: a review. Medical Physics 36(6Part1), 2052–2068 (2009)
14. Galloway, M.M.: Texture analysis using grey level run lengths. NASA STI/Recon Technical Report N 75 (1974)
15. Giger, M.L.: Machine learning in medical imaging. Journal of the American College of Radiology 15(3), 512–520 (2018)
16. Gorgel, P., Sertbas, A., Ucan, O.N.: A wavelet-based mammographic image denoising and enhancement with homomorphic filtering. Journal of Medical Systems 34(6), 993–1002 (2010)
17. Haas, B.M., Kalra, V., Geisel, J., Raghu, M., Durand, M., Philpotts, L.E.: Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. Radiology 269(3), 694–700 (2013)
18. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. pp. 610–621. No. 6, IEEE (1973)

19. Helvie, M., Roubidoux, M., Hadjiiski, L., Zhang, Y., Carson, P., Chan, H.: Tomosynthesis mammography versus conventional mammography: comparison of breast masses detection and characterization. In: Radiological Society of North America 93rd Scientific Assembly, Nov (2007)
20. Hinton, G.E.: Deep belief networks. Scholarpedia 4(5), 5947 (2009)
21. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
22. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.: Artificial intelligence in radiology. Nature Reviews Cancer 18(8), 500 (2018)
23. Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q.: Deep learning for image-based cancer detection and diagnosis- a survey. Pattern Recognition 83, 134–149 (2018)
24. Jamieson, A.R., Drukker, K., Giger, M.L.: Breast image feature learning with adaptive deconvolutional networks. In: Medical Imaging 2012: Computer-Aided Diagnosis. vol. 8315, p. 831506. International Society for Optics and Photonics (2012)
25. Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N.: Large scale deep learning for computer aided detection of mammographic lesions. Medical Image Analysis 35, 303–312 (2017)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
27. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436 (2015)
28. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Backpropagation applied to handwritten zip code recognition. Neural Comp. 1(4), 541–551 (1989)
29. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. vol. 86, pp. 2278–2324. Taipei, Taiwan (1998)
30. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision. pp. 21–37. Springer (2016)
31. Lu, L., Zheng, Y., Carneiro, G., Yang, L.: Deep learning and convolutional neural networks for medical image computing. Advances in Computer Vision and Pattern Recognition; Springer (2017)
32. Moura, D.C., López, M.A.G.: An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. International Journal of Computer Assisted Radiology and Surgery 8(4), 561–574 (2013)
33. Pak, F., Kanan, H.R., Alikhassi, A.: Breast cancer detection and classification in digital mammography based on non-subsampled contourlet transform (nsct) and super resolution. Computer Methods and Programs in Biomedicine 122(2), 89–107 (2015)
34. Petersen, K., Nielsen, M., Diao, P., Karssemeijer, N., Lillholm, M.: Breast tissue segmentation and mammographic risk scoring using deep learning. In: International Workshop on Digital Mammography. pp. 88–94. Springer (2014)
35. Poplack, S.P., Tosteson, T.D., Kogel, C.A., Nagy, H.M.: Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. American Journal of Roentgenology 189(3), 616–623 (2007)
36. Rojas-Domínguez, A., Nandi, A.K.: Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection. Computerized Medical Imaging and Graphics 32(4), 304–315 (2008)

37. Rojas-Domínguez, A., Nandi, A.K.: Development of tolerant features for characterization of masses in mammograms. Computers in Biology and Medicine 39(8), 678–688 (2009)
38. Rojas-Domínguez, A., Nandi, A.K.: Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. Pattern Recognition 42(6), 1138–1148 (2009)
39. Sahiner, B., Chan, H.P., Petrick, N., Helvie, M.A., Goodsitt, M.M.: Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. Medical Physics 25(4), 516–526 (1998)
40. Samala, R.K., Chan, H.P., Hadjiiski, L., Helvie, M.A., Wei, J., Cha, K.: Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Medical Physics 43(12), 6654–6666 (2016)
41. Sampat, M.P., Bovik, A.C., Whitman, G.J., Markey, M.K.: A model-based framework for the detection of spiculated masses on mammography a. Medical Physics 35(5), 2110–2123 (2008)
42. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual Review of Biomedical Engineering 19, 221–248 (2017)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
44. Sun, W., Tseng, T.L.B., Zhang, J., Qian, W.: Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Computerized Medical Imaging and Graphics 57, 4–9 (2017)
45. Suzuki, K.: Survey of deep learning applications to medical image analysis. Med Imaging Technol 35(4), 212–226 (2017)
46. Ting, F.F., Tan, Y.J., Sim, K.S.: Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications 120, 103–115 (2019)
47. Yanez-Vargas, I., Gonzalez-Reyna, S., Gonzalez-Ramirez, A., Guerrero-Gasca, I., Astudillo-Montenegro, F.: Super-resolution of mammograms based on analysis of wavelet family and iterative scales. In: In Proceeding International Conference on Electronics, Communications and Computers CONIELECOMP. pp. 1–5. IEEE (2017)

ISSN 1870-4069

Machine Learning Techniques for Diagnosis of Breast Cancer

Alfonso Rojas Domínguez

Tecnológico Nacional de México, Campus León, Mexico

alfonso.rojas@gmail.com

Abstract. Breast cancer is the most common form of cancer in the female population. As with any form of cancer, early detection of breast cancer is one of the most important factors affecting the possibility of recovery from the disease. Early detection of breast cancer can be achieved through mammography screening programs. Studies have shown that double reading of mammograms improves the detection of breast abnormalities. Unfortunately, the enormous amount of mammograms to be examined prohibits the practice of double reading by human experts. Computerized systems for automated detection and classification of breast abnormalities in mammograms have been developed as a possibility to alleviate this problem. These systems have the objective of accompanying human experts and eventually replace one of them in double or triple reading of mammograms. Some of the signs associated with breast cancer that can be observed in mammograms include: masses, calcifications, distortion of the parenchimal tissue, asymmetry of breast tissue between the breasts of a patient, etc. Beginning from about three decades ago, computerized systems and algorithms specialized in the detection and diagnosis of each type of these signs have been proposed by researchers and described in the scientific literature. Interest in the problem persists to this day; the development of automated systems for detection/diagnosis of breast cancer is currently a very active research area. In this work, an overview of the techniques developed for diagnosis of breast cancer is presented from a perspective of the work that has been preciously carried out personally in this area.

Keywords: machine learning, mammography analysis, breast cancer, CAD.

1 Introduction

Breast cancer is the most common form of cancer among women, and the second deadliest type of cancer (after lung cancer) in high-risk countries such as the United Kingdom and the United States [20]. Most medical experts agree that successful treatment of breast cancer is often linked to early diagnosis. Over 80% of patients diagnosed at the earliest stage of the disease are cured by

current therapies. Thus, while breast cancer in general cannot be prevented, it is clear that its detection at an early stage is one of the most important factors contributing to a positive prognosis. Physically, breast cancer manifests itself by one or more of the following breast abnormalities: micro calcifications, breast tumors (i.e. malignant breast masses), distortion of the breast tissue (known as architectural distortion), asymmetry of breast tissue between breasts of the same patient, etc.

All of these abnormalities can remain undetected for years even if monthly self-examination is performed, and many cases are completely asymptomatic under routine physical examination. Screening methods for breast cancer currently available or under development include: breast self-examination; mammography (and Digital Breast Tomosynthesis, DBT); magnetic resonance imaging (MRI); and breast ultrasound. Among these, mammography is the preferred screening method. This is due to its sensitivity (much higher than self-examination, although lower than MRI), specificity (higher than MRI), ability to reduce mortality (which self-examination does not possess) and other advantages such as relative cost (less expensive than MRI, DBT and ultrasound), relative non-invasiveness (MRI requires the injection of a chemical agent), etc.

A true-positive (TP) detection occurs when the screening indicates a positive prediction for a case that is later confirmed as cancer, and a false-positive (FP) detection occurs when the screening returns a positive prediction for a case that is later found to be normal. True-negatives (TN) and a false-negatives (FN) are similartly defined. The number of TPs, FPs, TNs and FNs in a test population can be combined into two measures that are used to summarize the efficiency of a screening method: sensitivity is defined as the number of TPs divided by the total amount of cancer cases in the test population; specificity is defined as the number of TNs divided by the total number of cases in the test population with no breast cancer. For example, a recent study on screening mammography reports an overall sensitivity of 86.6% and specificity of 96.8%, with a 0.6% incidence of breast cancer in the population of the study [3]. In other words, the particular screening program is almost always right when it returns a negative prediction (high specificity of $\approx 97\%$), but misses about 13% of the cancer cases (sensitivity is $\approx 87\%$).

Mammography is a specific type of imaging: the radiographic method of imaging the human breast using a low dose of x-ray radiation passed through the breast to form an image of its internal tissue. Nowadays there is a distinction between two slightly different methods of mammography: in conventional mammography the radiation exposes a photographic film. The images thus formed are known as mammograms, and can either be viewed directly on the film by means of a view box or on a work station with previous digitization of the image by electronic scanning. On the other hand, modern mammography systems make use of a digital image receptor instead of a photographic film to produce digital images directly; the method that use these systems is known as digital mammography, and the images that result from this method may sometimes be referred to as digital mammograms. Mammography can aid in the

early detection of breast cancers through the screening of asymptomatic women to detect occult abnormalities in the breast internal tissue. This modality is known as Screening Mammography. The first controlled trial of screening for breast cancer was performed in the 1960's [20,12]. A large amount of data is now available regarding screening mammography, however, many questions remain to be answered and some aspects of screening mammography (such as the particular efficacy in different age groups, or its cost-effectiveness), are still debated. It is generally agreed that mammographic screening reduces breast cancer mortality, and a number of leading health care organizations recommend screening mammography on an annual or two-year basis for all women over a certain age (some indicate a two-year mammogram from age 40 and annual screening from age 50, while others recommend annual screening from age 40), since the risk increases throughout a woman's lifetime.

Screening mammography is currently the most effective tool for early detection of breast cancer. However, the visual examination of a mammogram by a radiologist expert in search for abnormalities is a hard and time-consuming task because the images must be examined with great detail and attention. Furthermore, only approximately five out of a thousand cases examined will return findings related to breast cancer. As a result, radiologists fail to detect 10% to 30% of cancers [24]. On the other hand, the cancer detection rate increases about 5% if double reading of mammograms is employed in the screening process. Thus, even if the number of sufficiently trained radiologists were enough to perform human-based examination of the mammograms, the rate of false diagnosis would be predictively large given the difficulty of the task and the results obtained by current studies on the accuracy of diagnosis (positive predictive value is less than 35% [24]). Not surprisingly, the debate as to the best means for analyzing the large volume of screening mammograms has been focused on the possible use of computer technology, both to reduce the analysis time, as well as to increase the accuracy of diagnosis.

Since computer technology can be employed to aid in the detection and the diagnosis of breast cancer with mammography, two different types of systems are identified by researchers and developers: computer-aided detection (CADe) systems, and computer-aided diagnosis (CADx) systems. CADe systems have been developed to help radiologists in detecting lesions that may indicate the presence of breast cancer. Currently, the objective of these systems is to act as a second reader in double-reading of mammograms, where the final decision is made by the radiologist. The objective of CADx systems is to help radiologists in making a recommendation for patient management. CADx systems are used after a positive detection of a breast abnormality has occurred. If the abnormality is suspected to be malignant, a biopsy must be performed to confirm or reject this suspicion. Fig 1 shows a diagram of the typical components in a CADe/CADx system.

From the point of view of researchers, automated detection of micro calcifications is generally considered a well studied problem. This is not the case with other breast abnormalities such as masses and architectural distortions, where

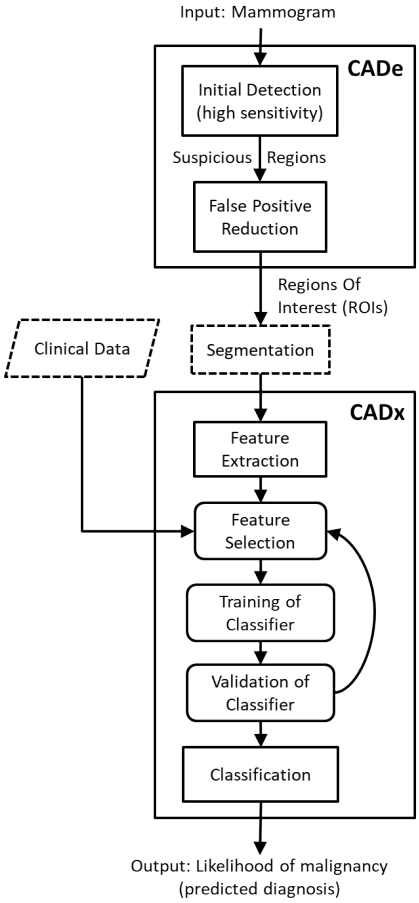


Fig. 1. Typical CADe and CADx systems. Rounded boxes represent stages that are exclusive to the design/development mode of the system. Solid-line square boxes represent stages corresponding to the operational mode of the system. Dashed-line boxes represent stages that may appear in the operational mode of the system or may not be present in that mode.

automated detection and diagnosis methods are still under development. The large variability in the appearance of breast masses, added to the significant overlap in the appearance of malignant and benign masses, and the fact that abnormalities (masses and other) are often occluded or hidden in dense breast tissue, make both their detection and diagnosis very difficult.

2 Diagnosis of Breast Cancer as a ML Task

Both of the computer-aided types of systems (CAD and CADx) for breast cancer with mammography are incarnations of the general model for a pattern

recognition system, with the incorporation of ad-hoc image processing and image analysis techniques.

Detection of breast masses involves the scanning of mammograms with the intention of identifying suspicious image regions that may correspond to breast masses. Automated detection methods aim to identify as many abnormalities as possible without labelling normal regions as suspicious. The variability of the appearance of abnormalities, coupled with the similarity between abnormalities and normal breast tissue, make detection a very difficult task.

The most common approach taken for the solution of this problem consists in three steps: 1) application of a mammogram-enhancement routine, 2) segmentation of all potentially suspicious regions, and 3) selection of the suspicious regions. The output of mass detection methods are all the regions labelled as suspicious, or Regions Of Interest (ROIs).

The segmentation of masses consists of separating the masses from the surrounding normal tissue in regions of interest. In this case, automated methods aim to produce a very precise segmentation, so that features that will be later used for the classification of abnormalities can be extracted with reliability.

One of the main obstacles associated with the automated segmentation of masses is the difficulty to determine whether a segmentation is correct or not. This is because in many cases the shape of breast masses is very complex, and the boundaries between mass and background tissue can appear obscured or undefined. Segmentation algorithms may be used as a pre-processing stage for classification of masses.

The ultimate objective of automated methods for classification of masses is to provide a tentative diagnosis (the final decision is produced by a human expert) of individual masses, based on their physical attributes. These methods are incarnations of a generic model for supervised pattern classification systems. According to this model, a classifier is presented with features obtained from a selection of the objects that are to be classified, in a process known as training. The trained classifier can later label objects which were not used in its training, an ability known as generalization. The performance of classification methods depends on the type and quality of the features employed to train the classifier. Features that possess some degree of robustness against segmentation errors are preferred.

The goal of CADx system is to perform the analysis of the ROIs returned by CADe systems. This analysis can be a pixel-based analysis or a region-based analysis. In a pixel-based analysis the features for the classifier are computed directly from every pixel that is part of the lesion in a particular ROI. In contrast, in a region-based analysis the features for the classifier are the result of an extra level of abstraction introduced by an intermediate representation between pixels and features.

3 Mass Detection Techniques

Although the detection of breast masses can be seen as an end on itself, in the context of Computer-Aided mammography-analysis it is often the first of a series of steps required to achieve the classification of abnormalities. Detection algorithms return the location of the breast masses in the mammograms; additionally, these can also return an approximate segmentation of the detected abnormalities, which can be used to perform the extraction of features required for automated classification.

In our previous work [21] we described a method for mammogram enhancement and abnormality detection. The proposed method is divided into three main stages. The first stage is a mammogram enhancement procedure that has the objective of improving the segmentation of the distinct structures in the mammogram when performed via simple conversion to binary images at multiple threshold levels. This enhancement algorithm is different from others in the literature in that it computes the parameter of the enhancing function in an adaptive manner, based on local statistics of the pixels in the mammographic image, and it considers multiple scales. The second stage consists of segmentation and feature extraction steps. In this stage, regions are segmented and several shape and gray-level characteristics of the regions are computed. Finally, in the third stage, a ranking system is employed to select suspicious regions. This is a novel approach to the problem of region selection (elimination of FPs) that does not require training, and implements a type of on-the-fly feature selection.

Our multi-thresholding segmentation method has been cited in different studies as a means to enable the extraction of classification features [6,7,2].

For instance, Al-Najdawi et al. published their own method for enhancement of mammograms and segmentation of breast masses, similar to what we described previously.

The authors also presented their method for classification of masses. Unfortunately, their classification method does not employ a robust classifier, but instead reaches a decision based on a set of heuristic rules and no validation or independent (from the training data) test results were reported. In contrast, Liu & Zeng [17] trained an Support Vector Machine (SVM) on texture features and geometry features, achieving a sensitivity of 76.9% at 1.43 FPs per image. In another example, Pak et al. published a CAD method in which the Non-Subsampled Contourlet Transform and Super Resolution are used for image enhancement, shape analysis features are extracted from thresholded regions and classified with AdaBoost (ANN, SVM and K-NN were also tested). They report a mean sensitivity of 87.15% with 93.58% specificity [19].

Finally, Choi et al. described their proposal of a method for the generation of a classifier ensemble for CADe/CADx systems on mammography [7]. The ensemble contained a number of base classifiers, each of which was associated with a particular feature set previously used in the literature. A total of twelve feature sets were considered, including our own proposal of features for characterization of spiculation and fuzziness of mass margin [22]. For the base classifiers, the authors compared experimentally between SVMs with radial basis function as

kernel and Neural Networks trained with backpropagation. Importantly, their study proved that an ensemble of classifiers can achieve a higher performance (for both the detection and the classification tasks) than single classifiers even if feature selection is used to boost the performance of the individual classifiers. Ensemble systems often overcome other strategies in many applications.

4 Robust Features for Breast Mass Classification

In our previous work [22,23] four new features for the analysis of breast masses were presented. These features were designed to be insensitive to the exact shape of the contour of the masses, so that an approximate contour, such as one extracted via an automated segmentation algorithm, can be employed in their computation. Two of the features, Sp_{SI} and Sp_{GO}, measure the degree of spiculation of a mass and its likelihood of being spiculated. The Sp_{GO} feature is a measure of the relative gradient orientation of pixels that correspond to possible spicules based on a feature known as Phase Congruence [16].

The Sp_{SI} feature is based on a comparison of mutual information measures between selected components of the mammographic images, which are obtained by means of Gabor filter banks. The last two features, Fz_1 and Fz_2, measure the local fuzziness of the mass margins based on points defined automatically. The features were tested for characterization (i.e. discrimination between circumscribed and spiculated masses) and diagnosis (i.e. discrimination between benign and malignant masses) of breast masses using a set of 319 masses and three different classifiers. In the characterization experiments the features produced a result of approximately 90% correct classification. In the diagnosis experiments, the performance achieved was approximately 76% of correct classification.

Similar ideas to the classification features described above have been published and have received high numbers of citations [18,1,14]. For instance, Casti et al. presented a multistage system for detection and classification of breast abnormalities [5].

The detection stage included analysis of a Gradient Vector Field, while the classification features were based on the response of multidirectional Gabor filters, clustering and differential feature extraction based on the clusters created. The system detected nearly 80% of the malignant tumors from the DDSM and MIAS datasets at 3.47 and 2.92 FPs per image, respectively.

Khan et al. also investigated the performance of different approaches for directional feature extraction for mass classification based on banks of Gabor filters [15]. The authors compared six approaches and concluded that a Windows based Statistical Magnitude Gabor Response method was significantly superior to the other approaches for the classification of masses vs. normal tissue and for the classification of benign vs. malignant masses. The classification was performed by means of an SVM adapted to deal with unbalanced datasets (since in this problem, normal instances are much more frequent than abnormal instances). Khan et al. also worked on finding optimized Gabor filter banks through an incremental clustering algorithm (for filter selection) and Particle

Swarm Optimization (to optimize the parameters of the filters). Their results indicate that optimization of the parameters of the banks produces significantly better results and comparable to those in the state-of-the-art.

In 2016, Jiao et al. presented a deep-feture based framework for breast masses classification [14]. Their framework employed a Convolutional Neural Network (CNN) to extract hierarchical features from different layers of the CNN. Then the authors introduced a decision mechanism with two classifiers and a similarity measure to produce a final result. The CNN was pretrained on the LSVRC dataset [8] and fine tuned using the DDSM. The classification results of this method were superior to those in the state-of-the-art for the DDSM. Similarly, other deep-based frameworks have been quite successful in solving this problem and are preferred nowadays over schemes based on the more "traditional" feature extraction methodologies [9,11,10]. Jiao et al. also described a parameter updating strategy for improving performance of pre-trained CNNs models (through a metric-learning network trained jointly with the CNN), further improving their classification results on the DDSM.

More recently, Al-antari et al. described a CADx system based on texture (statistical) features extrated from the Gray Level Co-occurrence Matrix (GLCM) [13] of breast abnormalities previously detected, together with invariant moments, and first order statistics for a total set of 347 features [1]. The authors compared the results of classification by means of LDA, QDA, NN and a Deep Belief Network (DBN), concluding that the DBN achieves significantly superior performance, although their experimentation employed a limited number of instances (220 masses, equally split between benign and malignant classes and 116 normal tissue samples).

5 Conclusion

A brief overview of techniques for computer-aided detection and diagnosis or breast abnormalities has been presented. The most important observations that can be made are the following:

1. There is a large variety of frameworks for CAD/CADx of breast cancer based on mammography analysis. Although many different features for breast masses detection and classification have been proposed over the years, the most popular and effective features are those based on robust texture (statistical) analysis and multidirectional/multiscale filtering for identification of spiculated masses.
2. Traditionally, researchers in this area have preferred the use of complex features followed by simple classifiers. Some of the classifiers that appear most often are LDA and (linear) SVMs. In some cases, to boost performance, ensembles of classifiers have been employed. Although the results can be significantly better than those of individual classifiers, the complexity of the system and the number of system parameters to be adjusted also increases. In some cases, a degree of over-fitting to the training data cannot be excluded.

3. From approximately 2015-to date, the deep learning models that have dominated in many pattern recognition applications, have been adopted by researchers working on CAD of breast cancer. However, due to the particular characteristics of the problem (such as unbalanced data, lack of sufficient amounts of labelled training data, and complexity of the problem), rather than using typical deep learning models, the preferred solutions include transfer learning from large generic datasets or from large amounts of handcrafted features, and hybrid/augmented systems that employ optimization methods and incorporate additional knowledge to improve the performance of regular deep models.

Ultimately, the goal of researchers should be to produce CAD systems that are as robust and directly applicable to different problem domains as possible. In our present case, this would mean to design CAD systems that could be used for different types of cancers, on different imaging modalities, not just mammogram analysis for breast cancer detection and diagnosis. Current research trends are moving in that direction, but at the same time, some new problems arise. For instance, although well-known methods for hyper-parameter optimization of learning models exist (i.e., evolutionary algorithms), training of deep models of sufficient scale is still very expensive computationally, so that the need for accelerated training algorithms exist. Similarly, although nowadays the techniques of dropout and data augmentation are successfully employed to ensure the generalization capabilities of deep models, steps for their systematic application are not yet well established. Clearly, these challenges are not exclusive of CAD systems, but the problems of automated breast cancer detection and diagnosis, because of their intrinsic characteristics, lend themselves to the study of interesting solutions to said challenges.

Finally, regarding the clinical applicability of CAD systems, a recent study indicates that the implementation of CAD as part of the clinical practice requires initial training and involves a learning curve at the beginning of which (first two months) the recall rates may increase dramatically [4]. The same study also reports that some radiologists find the marks produced by CAD systems annoying (there are too many FPs) or threatening (because the use of CAD may be interpreted as if the radiologists were incompetent). Nevertheless, most of the current evidence supports the conclusion that CAD for screening mammography increases the detection sensitivity with a reasonable increase in the recall rate.

List of Acronyms and Abbreviations.

– ANN: Artificial Neural Network.
– CAD(e)/CADx: Computer Aided Detection/Diagnosis.
– CNN: Convolutional Neural Network.
– DBN: Deep Belief Network.
– DBT: Digital Breast Tomosynthesis.
– DDSM: Digital Database for Screening Mammography.
– FN/FP: False Negative/Positive.
– GLCM: Gray Level Co-occurrence Matrix.

– K-NN: K-Nearest Neighbors.
– LDA: Linear Discriminant Analysis.
– LSVRC: Large Scale Visual Recognition Competition.
– MIAS: Mammographic Image Analysis Society.
– ML: Machine Learning.
– MRI: Magnetic Resonance Imaging.
– QDA: Quadratic Discriminant Analysis.
– ROI: Region of Interest.
– SVM: Support Vector Machine.
– TN/TP: True Negative/Positive.

Acknowledgments. This work was supported by the National Council of Science and Technology of Mexico (CONACYT) under Research Grant CÁTEDRAS-2598 (A. Rojas).

References

1. Al-antari, M.A., Al-masni, M.A., Park, S.U., Park, J., Metwally, M.K., Kadah, Y.M., Han, S.M., Kim, T.S.: An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network. Journal of Medical and Biological Engineering 38(3), 443–456 (2018)
2. Al-Najdawi, N., Biltawi, M., Tedmori, S.: Mammogram image visual enhancement, mass segmentation and classification. Applied Soft Computing 35, 175–185 (2015)
3. Banks, E., Reeves, G., Beral, V., Bull, D., Crossley, B., Simmonds, M., Hilton, E., Bailey, S., Barrett, N., Briers, P., et al.: Influence of personal characteristics of individual women on sensitivity and specificity of mammography in the million women study: cohort study. BMJ 329(7464), 477 (2004)
4. Birdwell, R.L.: The preponderance of evidence supports computer-aided detection for screening mammography. Radiology 253(1), 9–16 (2009)
5. Casti, P., Mencattini, A., Salmeri, M., Ancona, A., Mangeri, F., Pepe, M.L., Rangayyan, R.M.: Contour-independent detection and classification of mammographic lesions. Biomedical Signal Processing and Control 25, 165–177 (2016)
6. Chaddad, A.: Automated feature extraction in brain tumor by magnetic resonance imaging using gaussian mixture models. Journal of Biomedical Imaging 2015, 8 (2015)
7. Choi, J.Y., Kim, D.H., Plataniotis, K.N., Ro, Y.M.: Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography. Expert Systems with Applications 46, 106–121 (2016)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
9. Dhungel, N., Carneiro, G., Bradley, A.P.: Automated mass detection in mammograms using cascaded deep learning and random forests. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8 (2015)

10. Dhungel, N., Carneiro, G., Bradley, A.P.: A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Medical Image Analysis 37, 114–128 (2017)
11. Dhungel, N., Carneiro, G., Bradley, A.P.: Fully automated classification of mammograms using deep residual neural networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 310–314. IEEE (2017)
12. Dixon, J.M.: Breast cancer: diagnosis and management. Elsevier Science Health Science div (2000)
13. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 6, 610–621 (1973)
14. Jiao, Z., Gao, X., Wang, Y., Li, J.: A deep feature based framework for breast masses classification. Neurocomputing 197, 221–231 (2016)
15. Khan, S., Hussain, M., Aboalsamh, H., Bebis, G.: A comparison of different gabor feature extraction approaches for mass classification in mammography. Multimedia Tools and Applications 76(1), 33–57 (2017)
16. Kovesi, P., et al.: Image features from phase congruency. Videre: Journal of Computer Vision Research 1(3), 1–26 (1999)
17. Liu, X., Zeng, Z.: A new automatic mass detection method for breast cancer with false positive reduction. Neurocomputing 152, 388–402 (2015)
18. de Oliveira, F.S.S., de Carvalho Filho, A.O., Silva, A.C., de Paiva, A.C., Gattass, M.: Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and svm. Computers in Biology and Medicine 57, 42–53 (2015)
19. Pak, F., Kanan, H.R., Alikhassi, A.: Breast cancer detection and classification in digital mammography based on non-subsampled contourlet transform (nsct) and super resolution. Computer Methods and Programs in Biomedicine 122(2), 89–107 (2015)
20. Patlak, M., Nass, S.J., Henderson, I.C., Lashof, J.: Mammography and beyond: developing technologies for the early detection of breast cancer, vol. 2. National Academy Press (2001)
21. Rojas-Domínguez, A., Nandi, A.K.: Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection. Computerized Medical Imaging and Graphics 32(4), 304–315 (2008)
22. Rojas-Domínguez, A., Nandi, A.K.: Development of tolerant features for characterization of masses in mammograms. Computers in Biology and Medicine 39(8), 678–688 (2009)
23. Rojas-Domínguez, A., Nandi, A.K.: Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. Pattern Recognition 42(6), 1138–1148 (2009)
24. Sampat, M.P., Markey, M.K., Bovik, A.C., et al.: Computer-aided detection and diagnosis in mammography. Handbook of Image and Video Processing 2(1), 1195–1217 (2005)

ISSN 1870-4069

Evaluation of Breast Cancer by Infrared Thermography

Antony Morales Cervantes[1] , Eleazar Samuel Kolosovas Machuca[2],
Edgar Guevara[2], Francisco Javier González[2], Juan J. Flores[1]

[1] Universidad Michoacana de San Nicolás de Hidalgo,
Facultad de Ingeniería Eléctrica,
Mexico

[2] Universidad Autónoma de San Luis Potosí,
Coordinación para la Innovación y
la Aplicación de la Ciencia y la Tecnología,
Mexico

juanf@umich.mx

Abstract. Breast cancer is one of the leading causes of death in women. Temperature measurement by means of thermography has several advantages; It is non-invasive, non-destructive and it is profitable. The measurement of breast temperature by infrared thermography is useful for detecting changes in blood perfusion that may occur due to inflammation, angiogenesis or other pathological causes. In this work, 206 thermograms of patients with suspected breast cancer were analyzed, using a classification method, in which thermal asymmetries were calculated. The most vascularized areas of each breast were extracted and compared; these two metrics were then added to obtain a thermal score, indicative of thermal anomalies. The classification method based on this thermal score allowed us to evaluate the effectiveness of the test, obtaining a sensitivity of 100%, specificity of 68.68%; a positive predictive value of 11.42% and a negative predictive value of 100%. These results highlight the potential of using infrared thermography as a complementary tool to mammography in the detection of breast cancer.

Keywords: breast cancer, infrared thermography, image analysis.

1 Introduction

Breast cancer is one of the leading causes of death in women in recent years [1, 2], affects all social levels of the population, is the first most common cancer in Mexico with 18.7%, followed by cancer of digestive organs with 18.0% [3]. The likelihood of a person developing breast cancer depends on some factors that unfortunately cannot be avoided. These include age, sex and genetics. However, there are some typical characteristics of the presence of cancer, such as tumours and specific tissue activity that specialists have been using to make an early diagnosis [4].

Studies have shown that early detection of cancer ensures a better prognosis and is essential to have a higher survival rate, if detected early, the cure rate is 95% [5].

Breast imaging techniques have been developed as primary clinical methods for the identification of early and differentiated breast cancers of benign breast tumours [6]. Mammography is the most common imaging technique used for breast cancer screening. However, the false-negative rate can reach up to 30% and exposes patients to ionizing radiation [7]. In addition, mammography is less effective in young women and in those who have dense breast tissue [8]. Ultrasound is mainly used to differentiate the properties of solid and cystic breast lesions identified by mammography. Dense breast tissue can be examined by aspiration-guided biopsy and preoperative localization. Due to the time required to perform an exam, the need for proper management training and other limitations, ultrasound alone is not suitable as a screening method for breast cancer. In fact, ultrasound and mammography can ignore many cases in which the tumour is <0.5 cm [6].

In the 1960s, infrared thermography began to be used in medical diagnosis, but until 1982 it was approved by the Food and Drug Administration (FDA) as a complementary tool for breast cancer detection [9]. Since then, the sensitivity of infrared imaging technology has increased substantially and has become a more powerful tool for the diagnosis of breast cancer [10]. The measurement of temperature through infrared thermography is advantageous since it is completely non-invasive, non-destructive, cost-effective and can provide temperature data that give a distribution over a wide area [11]. The thermal analysis of the skin's temperature distribution in order to obtain information about a possible internal tumour offers more advantages to indicate an abnormal metabolism in the early stages of cancer [12]. Therefore, thermography is very convenient for locating changes in blood perfusion that may occur due to inflammation, angiogenesis or other causes. It is known that asymmetric temperature distributions, as well as the presence of hot and cold temperature points, are good indicators of an underlying problem [13].

Although the diagnosis is usually carried out manually by experts, there is a high demand for automatic methods that can also be used as a second opinion [14]. Automated thermogram analysis consists of dividing the image into segments of interest and analyzing it later Image segmentation refers to the technique that divides a digital image into multiple sections and is generally used to identify regions of interest or other relevant information in digital images [15]. One of the main ways to differentiate anomalies in the breasts is the comparison by thermal asymmetries in which the left breast is compared with the right breast. Another important point that has been evaluated is the difference in temperature that can exist in both breasts. In this work, an effective approach to automatically analyze breast thermograms for cancer diagnosis is presented.

1.1 Interpretation of Images

The first methods of interpretation of infrared images of breast were based solely on qualitative criteria (subjective). The images were read to see the variations in the vascular pattern without taking into account the temperature variations between the breasts (Tricore method) [16]. This resulted in wide variations in the results of studies conducted with inexperienced interpreters.

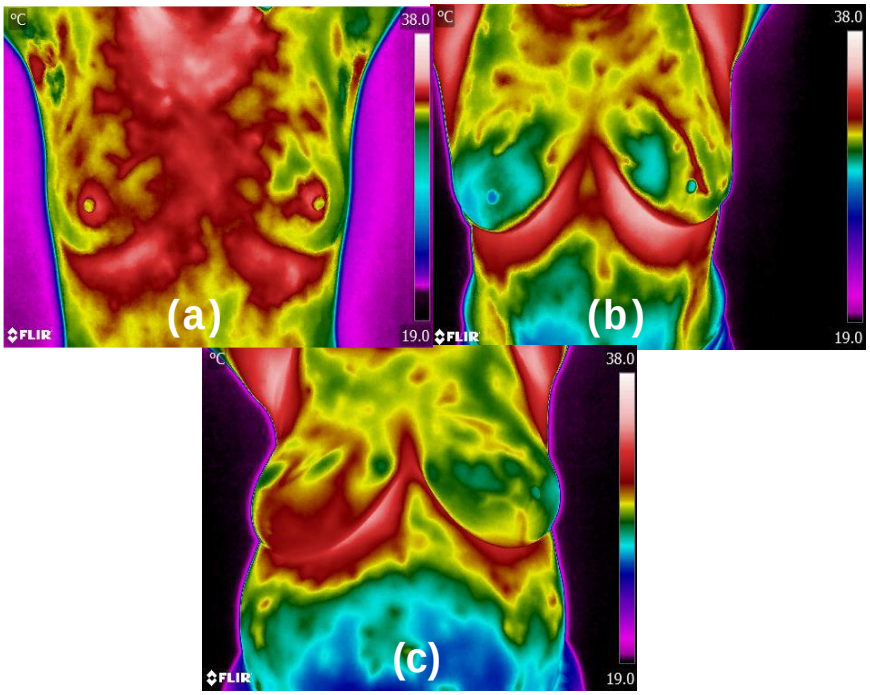


Fig. 1. Different types of vascularity (a) TH2 (normal uniform), (b) TH4 (abnormal), and (c) TH5 (severely abnormal).

Research throughout the 1970s showed that when both qualitative and quantitative data were incorporated into interpretations, an increase in sensitivity and specificity was performed. In the early 1980s, a standardized method of thermovascular analysis was proposed. The interpretation is composed of 20 discrete vascular and breast temperature attributes [17, 18].

1.1.1 Thermobiological Classification

This method of analysis was based on previous research and large-scale studies that included tens of thousands of patients. Using this methodology, thermograms are rated in one of the 5 TH (thermobiological) classifications. Based on the combined vascular pattern and temperatures across the two breasts, the images would be classified as TH1 (normal non-vascular uniform), TH2 (normal uniform vascular), TH3 (equivocal), TH4 (abnormal) or TH5 (severely abnormal) (see Fig. 1).

The use of this standardized interpretation method significantly increased the sensitivity, specificity, positive and negative predictive value of infrared images and the reliability of interpretation. The patient's continuous observations and investigations during the last two decades have caused changes in some of the thermovascular values; therefore, keeping the interpretation system updated.

Variations in this methodology have also been adopted with great success. However, it is recognized that, as with any other imaging procedure, specialized training and experience produce the highest level of success in screening.

1.1.2 Ville Marie Infrared Graduation Scale

Ths rating scale is based on relevant clinical information when comparing infrared images of both breasts and current images with previous images. An abnormal infrared image requires the presence of at least one abnormal sign (Table 1).

Table 1. Ville Marie infrared (IR) graduation scale.

Abnormal signs
1. Significant vascular asymmetry.
2. Vascular anarchy consisting of unusual tortuous or serpiginous vessels that form clusters, loops, abnormal tree planting or aberrant patterns.
3. A temperature rise of 1 ° C on the scale (DT) when compared with the contralateral site and when associated with the area of clinical abnormality.
4. A focal DT of 2 ° C against the contralateral site.
5. A focal DT of 3 ° C against the rest of the ipsilateral breast when it is not present on the contralateral side.
6. Global DT of the sinuses of 1.5 ° C against the contralateral sinus.
Infrared scale
IR1 = From the absence of any vascular pattern to mild vascular asymmetry.
IR2 = From significant but symmetrical vascular pattern to vascular asymmetry moderate, particularly if it is stable.
IR3 = An abnormal sign.
IR4 = Two abnormal signs.
IR5 = Three abnormal signs.

2 Materials and Methods

The sample consists of 206 patients from the "Dr Raymundo Abarca Alarcón" General Hospital in Chilpancingo, Guerrero, Mexico. Ages between 17 and 74 years. Average age: 42.4 years with a standard deviation of 10.4, average body mass index of 27.8 with a standard deviation of 4.8 and without dermatological diseases.

The study was presented and approved by the hospital's ethics committee. Patients invited to participate in the study had clinical evidence of a tumour suggestive of cancer, risk factors for breast cancer and went to the clinic. None of the patients had declared cancer at the time of inviting them to participate in the study, first the thermographic image was taken and then the mammogram scheduled for evaluation was performed. Patients read the informed consent form before signing it.

The specialists performed the BI-RADS classification of mammography, clinical diagnosis and biopsy were performed on those who had a suspicious anomaly in the evaluation. The camera used was the IR FlexCam Pro R, with a focal plane matrix (FPA) detector, based on vanadium oxide (VOX) uncooled microbolometer, thermal sensitivity @ 30Hz: = 0.070 ° C at 30 C, temperature range -20 ° C to 100 ° C, ± 2%

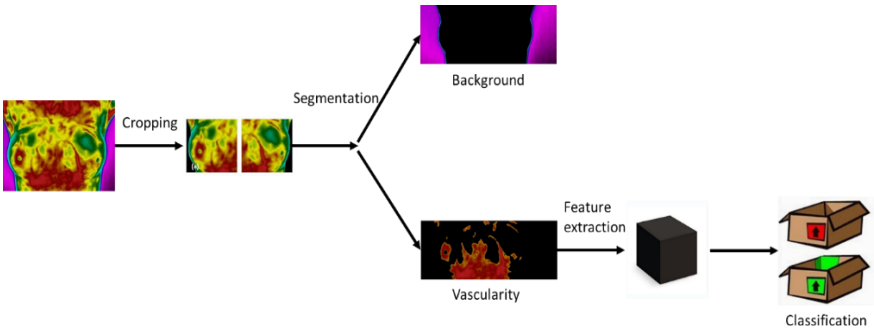


Fig. 2. Flowchart of the image processing performed.

Table 2. Scale of vascularitation.

(1) Absence of vascular patterns.
(2) Symmetric or moderate vascular patterns were found.
(3) Significant vascular asymmetry.
(4) Extended vascular asymmetry in at least one-third of the sinus area.

accuracy and a 20 mm f / 0.8 Germanium lens with a 23 ° horizontal x 17 ° vertical field of view.

The emissivity was set at 0.97 [15]. Patients who participated in the acquisition of thermal imaging did not perform physical activities, drank alcohol, smoked or used deodorant during the day the images were taken. At the time the image was taken, an acclimatization process of the patient was carried out in which they were asked to remain naked from the waist up for 20 minutes in a room with a controlled temperature of 24 ± 1 ° C.

Direct airflow to the patient was avoided and there were no nearby instruments that emit heat. The thermographic images were taken standing, with the hands holding the neck, 1.5 m from the camera, 5 photos were taken in total, one frontal, left lateral, right lateral and both frontal sinuses separately.

The automated program developed in this work is based on Gonzalez's work [10], which performs a simulation of a breast and a cancer tumor to evaluate desease through thermographic images. The interpretation of the image is done by means of a thermal score derived from the scale of the Ville Marie infrared classification [19]. This thermal score takes into account the two most significant infrared data that are: (a) the difference in surface temperature in the lesion compared to the specular imaging site in the contralateral breast (DT), and (b) the vascular pattern around and at the site of the injury [20].

The thermal score is calculated by adding the amount of vascularization to the difference in surface temperature in degrees Celsius at the site of the lesion compared to that of the contralateral breast. The amount of vascularization is determined using

the scale shown in Table 2. Fig. 2 shows the flow chart of the processing used in this work.

2.1 Left and Right Breast Segmentation

Image segmentation refers to the technique that divides a digital image into multiple segments. Segmentation is used to identify regions of interest or other relevant information in digital images [15].

The method used for the segmentation of the left and right breast from the rest of the body is based on the analysis of the projection profile [21]. It is used to find the upper, lower, left and right edges of the detected edge of the breast thermographic image. The Horizontal (or vertical) projection profile is a histogram of a matrix with a number of entries equal to the number of rows (or columns). The number of black pixels or white pixels in a row (or column) is stored in the corresponding entry.

The Horizontal Projection Profile (HPP) is used to locate the upper and lower edges. The Vertical Projection Profile (VPP) is used to find the left and right edges. First, the breast thermographic image becomes a grayscale image. Then, the following sequences of operations are performed on the image: Image filtering, edge detection, lower edge detection, upper edge detection, image threshold, left and right edge detection, central axis location, and segmentation of the left and right breast.

2.2 Edge Detection with the Sobel Operator

The Sobel operator measures the 2 D spatial gradient of the thermographic image and emphasizes the regions of high spatial frequency that correspond to the edges. The operator consists of a pair of 3 x 3 kernels to perform the convolution as shown in Eq. (4). One mask is simply the other turned 90°. These masks are designed to enhance the edges vertically and horizontally.

The Sobel operator gives a smoothing effect (average filter) and reduces false edges. In theory, the image gradient f (x, y) is a vector and is given by:

$$\nabla = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}, \tag{1}$$

where ∇ is the gradient operator. The magnitude of the gradient is given by:

$$Mag(\nabla f) = \sqrt{G_x^2 + G_y^2}\,, \tag{2}$$

where Mag (∇f) gives the magnitude of the edge to a particular location x-y.

The direction of the edge is found by:

$$\alpha(x, y) = tan^{-1}\frac{G_y}{G_x}. \tag{3}$$

The coefficient matrix for the Sobel operator is defined as:

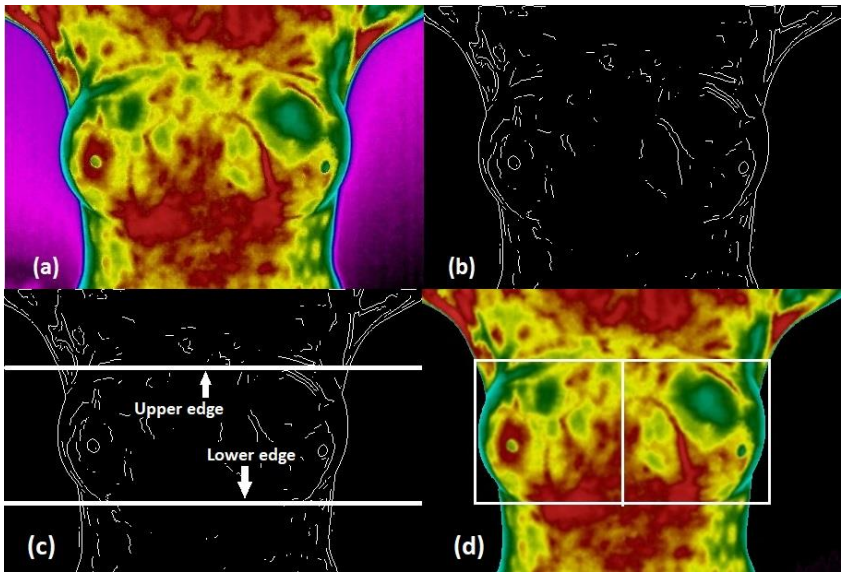


Fig. 3. Image segmentation process. (a) original image, (b) binarized image with edge detection, (c) upper and lower limits detected and (d) upper, lower, left and right edges detected.

$$H_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, H_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 2 \\ 1 & 2 & 1 \end{bmatrix}. \quad (4)$$

The filter results produce local gradient estimates for all pixels in the image in their two different directions, maintaining the following relation:

$$\nabla I(x,y) \approx \frac{1}{8}\begin{bmatrix} H_x \cdot I \\ H_y \cdot I \end{bmatrix}. \quad (5)$$

The result of the filters for each of the different senses is given by:

$$D_x(x,y) = H_x * I, D_y(x,y) = H_y * I. \quad (6)$$

Fig. 3 shows the results of edge detection of the breast thermographic image. The limit of the breast thermographic image is successfully detected using the Sobel operator as shown in Fig. 3 b).

2.3 Detection of the Upper and Lower Part of the Breasts

To find the lower breast, the edges detected in the image are scanned horizontally, to count the number of white pixels in each row of the bottom of the image.

In the infra-mammary line, the number of white pixels increases due to the infra-mammary fold of the breast. The scan is repeated until an HPP value equal to or greater than a threshold value is obtained. The HPP value will be smaller below the infra-mammary edge of the breast, as shown in Fig. 3 b). The row number corresponding to

the first high HPP value is taken as the lower limit (LL) for segmentation of the breast thermographic image.

HPP is used again to find the armpit location, which is considered as the upper end of the breast. Breast height normalization is necessary due to the variable height of the images. The breast thermographic image height varies depending on the structure and size of the breast. To standardize the height of the image, the distance between the lower edge detected and the lower part of the image is measured. The distance value varies depending on the structure and size of the breast. One study found that the distance value will be high for small breasts and less for large breasts [1]. According to the study and the observation, the height of the breast is calculated as indicated below.

1. If the distance between the bottom of the image and the lower limit of the breast is less than 26.3%[1] total pixels of the height then:

$$h = \frac{1}{2}m\,, \tag{7}$$

where h is the height of the area of interest of the image and m is the total number of rows present in the image.

2. If the distance between the bottom of the image and the lower limit of the breast is greater than 26.3% total pixels of the height then:

$$h = \frac{5}{6}m\,. \tag{8}$$

3. The upper limit (LS) shall be located in the row position given by:

$$LS = LI - h\,. \tag{9}$$

Finally, the upper and lower limits detected are shown in Fig. 3 (c). Horizontal lines along the upper and lower limits were drawn in the images to illustrate the process, but they do not represent real data.

2.4 Right and Left Edge Detection

After the thermographic image of the breast is segmented from the unwanted upper and lower part, the left and right edges are detected from the image using the vertical projection profile (PPV) method. This algorithm (PPV) is defined as the white pixel count number for each column. The steps followed for the detection of the left border are given below:

1. Segment the image with the upper and lower edges and apply Sobel.

2. To find the left limit, the image is analyzed from right to left and, if a white pixel is found, the position of the last column where it was found is stored, pointing only to the left-most white pixel.

[1] This percentage was determined empiricaly to capture and crop the breast area [13].

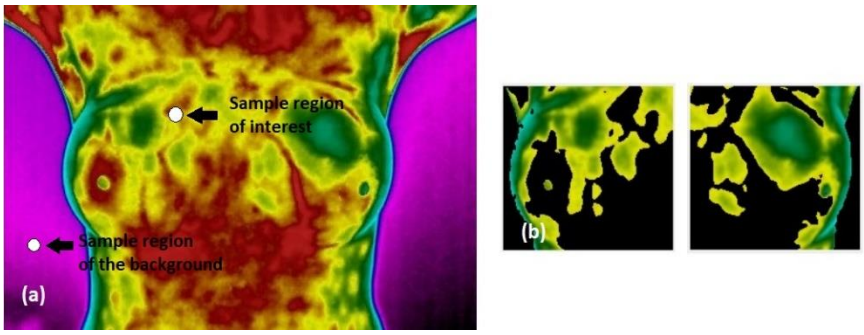


Fig. 4. Vascular areas segmentation. (a) Regions samples used to find the vascularity and eliminate the background in the images and (b) Segmented sinuses with vascularized area removed.

3. To find the right limit, the image is analyzed from left to right and, if a white pixel is found, the position of the last column where it was found is stored, pointing only to the white pixel farthest to the right. Therefore, the left and right edges detected are used to form a thermographic image of the desired breast by removing the unwanted left and right parts. In addition, the central axis of the breast is determined by dividing the width of the new image divided by two. Finally, Fig. 3(d) shows a thermographic image of the breast with the edges highlighted at the bottom, top, left and right.

2.5 Thermal Vascularity

An important part of image processing is the calculation of thermal vascularization. This process is performed in the color space known as CIELAB, which is normally used to describe all the colors that the human eye can perceive. In fact, it is done by removing the reddest parts of the image.

Fig. 3 (d) shows the areas with the highest temperature. The color space L * a * b * (also known as CIELAB or CIE L * a * b *) allows quantification of visual differences. In this color model, space is defined by three variables: L * represents brightness, and a * and b * correspond to the hue components. a * defines the distance along the red-green axis, and b * along the blue-yellow axis, these axes define the CIEXYZ space [22].

During the calculation of thermal vascularization, first, an image in the RGB space is converted to the L * a * b * space. Then, two sample regions of interest are selected; one in the vascular area and one in the background. We obtain the average a * b * of the selected areas, as shown in Figure 3.4 a). These values serve as markers for space a * and b * of background and vascular pixels. Next, each pixel is classified by calculating the Euclidean distance between that pixel and the marker. If that distance is too small, the pixel will be labelled as the closest marker. The result is a binary array that indicates each pixel's class.

Subsequently, the masks created in the background and red regions are used to segment the image by color, as well as to find vascularization. Once the original image is segmented by colors, the same values (upper, lower, left, right, and center) are used to separate the contours of the breasts and vascularized areas in order to calculate the

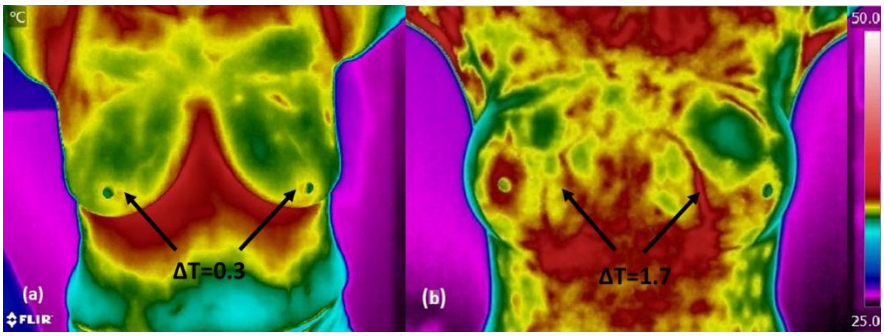


Fig. 5. Cancer diagnosis process. (a) IR image of healthy breasts. (DT) represents the difference in the surface of the breast. The calculated thermal score was found to be 1.3, obtained by adding the amount of vascularity (1): Absence of vascular patterns, while a value of 0.3 corresponds to the difference in surface temperature, DT, at the location of the lesion, compared to the contralateral breast and (b) IR image of a patient with infiltrating ductal carcinoma. The calculated thermal score was assigned to 3.7, obtained by adding the amount of vascularity (2): the two values denote symmetrical or moderate vascular patterns and a value of 1.7, associated with the surface temperature difference, DT, at the site of the lesion compared to the contralateral breast.

thermal score (Fig. 4 b). Subsequently, values from 1 to 4 are assigned according to the scale shown in Table 2.

Finally, the temperature difference of the heat source located in one of the breasts is measured with respect to its contralateral part. This is done with the thermal camera in real-time, that is, taking the images, observing them in one breast and then in their counterpart to obtain the temperature difference of interest. Then, a delta temperature is added to the thermal score (with a value of 2) of the vascularization found above, as shown in Fig. 5 b).

3 Results

All patients with a BI-RADS indicating the possibility of cancer underwent a biopsy. Of those patients, 8 of them exhibited infiltrating ductal carcinoma. Thermograms were analyzed using thermal scoring as presented in the previous section. Here, the infrared images were divided into two groups, (1) those with a thermal score below 2.5 were classified as healthy, (2) those with a thermal score greater than or equal to 2.5 were classified with some anomaly. Once the classification is performed on all images, a final score is assigned to the corresponding thermal image, for example, Fig. 3.5 b) shows a cancer patient with a thermal score of 6.7.

Further analysis of the thermographic images reveals statistical data as follows: of 206 patients, 8 true positives and 62 false positives were found. In addition, 136 were classified as true negatives and there were no false negatives. Obtaining a sensitivity of 100% with a specificity of 68.68%, a positive predictive value of 11.42%, and a negative predictive value of 100%. Fig. 6 shows the receiver operating characteristics (ROC) curve of thermographic analyses using the automated program. Based on the results, the ROC curve allows us to infer that it is possible to increase the specificity

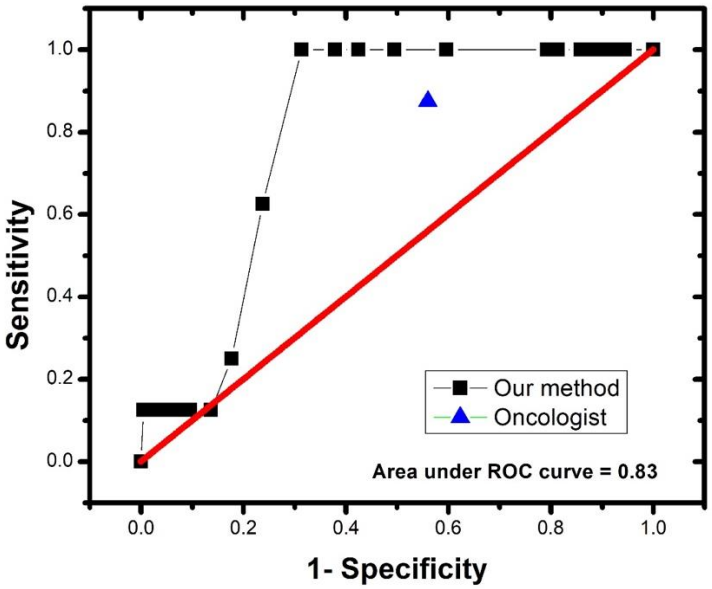


Fig. 6. ROC curve of the automated thermogram classification program.

Table 3. Comparison between our automated method and qualitative evaluation by an oncologist expert in thermography.

	Oncologist	**Our program**
Total of patients	206	206
Sick patients	8	8
Healthy patients	198	198
True positive	7	8
False positives	87	62
True negatives	111	136
False negatives	1	0
Sensitivity	87.50%	100%
Specificity	56.06%	68.68%
VPP	7.44%	11.42%
VPN	99.10%	100%

with this method by moving the cut-off point of the test, otherwise, the sensitivity would decrease considerably.

It is worth mentioning that the same thermographic images were analyzed qualitatively by an oncologist in a double-blind study, their findings were 7 true positives and 87 false positives, while 111 were classified as true negative and 1 false negative. The comparison of the results of the qualitative and quantitative method with our method is shown in Table 3.

4 Conclusions

A classification of thermographic images for the detection of breast cancer was performed using an automated program. 206 patients were considered for the screening test with clinical evidence of a tumour risk factor for breast cancer, the BI-RADS classification of mammography, the clinical diagnosis and the pathological results of the biopsy. The analyzed thermograms were classified as healthy (<2.5 thermal scores) or with an anomaly (≥2.5 thermal scores). The findings revealed that patients classified as healthy really had a healthy state (automated program) with a sensitivity of 100% and a specificity of 68.68%. In contrast, the same images qualitatively analyzed by an expert showed a sensitivity of around 87.5% and a specificity at 56%, so our results showed a significant improvement over a manual procedure.

In addition, an automated method to analyze thermograms was implemented, increasing the sensitivity and specificity of the test under study. The main objective will be to help the experts helping them with a better detection tool or even providing the possibility that someone without experience can benefit from the test results. We can emphasize that infrared thermography is not intended to replace mammography, but it is an excellent primary method/technique before patients undergo X-rays. It could be considered as a complementary diagnostic method to improve breast cancer detection.

Mamographies are an invasive and painful procedure. The proposed method is not intended to replace mamographies, but to avoid the suffering of undergoing one when it is absolutely not necessary. Using the proposed methodology, a mammography can be applied only those patients whose thermographic analysis indicates a breast abnormality.

References

1. Rastghalam, R., Pourghassem, H.: Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images. Pattern Recognition, 51, pp. 176–186 (2016)
2. González, F.: Thermal simulation of breast tumours. Revista Mexicana de Física, 53(4), pp. 323–326 (2007)
3. INEGI: Instituto Nacional de Estadística y Geografía (2015)
4. Guzman-Cabrera, R., Guzman-Sepulveda, J.R., Parada, A.G., Garcia, J.R., Cisneros, M.T., Baleanu, D.: Digital processing of thermographic images for medical applications. Revista De Chimie, 67(1), pp. 53–56 (2016)
5. Gautherie, M.: Thermopathology of breast cancer: measurement and analysis. Annals of the New York Academy of Sciences, pp. 383–415 (1980)
6. Yao, X., Wei, W., Li, J., Wang, L., Xu, Z.L., Wan, Y., Li, K., Sun, S.: A comparison of mammography, ultrasonography, and far-infrared thermography with pathological results in screening and early diagnosis of breast cancer. Asian Biomedicine, 8(1), pp. 11–19 (2014)
7. Boquete, L., Ortega, S., Miguel-Jiménez, J.M., Rodríguez-Ascaris, J.M., Blanco, R.: Automated detection of breast cancer in thermal infrared images, based on independent component analysis. Journal of Medical Systems, 36(1), pp. 103–111 (2012)
8. Wishart, G.C., Campisi, M., Boswell, M., Chapman, D., Shackleton, V., Iddles, S., Hallett, A., Britton, P.D.: The accuracy of digital infrared imaging for breast cancer detection in

women undergoing breast biopsy. European Journal of Surgical Oncology, 36(6), pp. 535–540 (2010)
9. Arora, N., Martins, D., Ruggerio, D., Tousimis, E., Swistel, A.J., Osborne, M.P., Simmons, R.M.: Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer. American Journal of Surgery, 196(4), pp. 523–526 (2008)
10. González, F.J.: Non-invasive estimation of the metabolic heat production of breast tumours using digital infrared imaging. Quantitative InfraRed Thermography Journal, 8(2), pp. 139–148 (2011)
11. Han, F., Shi, G., Liang, C., Wang, L., Li, K.: A simple and efficient method for breast cancer diagnosis based on infrared thermal imaging. Cell Biochemistry and Biophysics, 71(1), pp. 491–498 (2014)
12. Kathryn, J.C., Sireesha, G.V., Stanley, L.: Triple negative breast cancer cell lines: One tool in the search for better treatment of triple negative breast cancer. Breast Dis, 32, pp. 35–48 (2012)
13. Schaefer, G.: ACO classification of thermogram symmetry features for breast cancer diagnosis. Memetic Computing, 6(3), pp. 207–212 (2014)
14. Krawczyk, B., Schaefer, G.: A hybrid classifier committee for analysing asymmetry features in breast thermograms. Applied Soft Computing Journal, 20, pp. 112–118 (2014)
15. Zhang, X., Li, X., Feng, Y.: A medical image segmentation algorithm based on bi-directional region growing. Optik, 126(20), pp. 2398–2404 (2015)
16. Gauthrie, M., Kotewicz, A., Gueblez, P.: Accurate and objective evaluation of breast thermograms: basic principles and new advances with special reference to an improved computer-assisted scoring system. Thermal Assessment of Breast Health, pp. 72–93 (1983)
17. Hobbins, W.B.: Abnormal thermogram—significance in breast cancer. Interamer. J. Rad, 12, pp. 337 (1987)
18. Gautherie, M.: New protocol for the evaluation of breast thermograms. Thermological Methods, pp. 227–235 (1985)
19. Keyserlingk, J.R., Ahlgren, P.D., Yu, E., Belliveau, N.: Infrared imaging of the breast: initial reappraisal using high-resolution digital technology in 100 successive cases of stage i and ii breast cancer. pp. 245–251 (1998)
20. Wang, J., Chang, K.J., Chen, C.Y., Chien, K.L., Tsai, Y.S., Wu, Y.M., Teng, Y.C., Shih, T.: Evaluation of the diagnostic performance of infrared imaging of the breast: a preliminary study. BioMedical Engineering OnLine, 9(1), pp. 3 (2010)
21. Dayakshini, D., Kamath, S., Prasad, K., Rajagopal, K.V.: Segmentation of breast thermogram images for the detection of breast cancer – a projection profile approach. Journal of Image and Graphics, 3(1), pp. 47–51 (2015)
22. Cuevas, E., Zaldívar, D., Pérez, M.: Procesamiento digital de imágenes con MATLAB & Simulink. Ra-Ma (2016)

ISSN 1870-4069

Cancer Metastasis and the Immune System Response: CM-IS Modeling by Ising Model

Matias Alvarado[1], Renato Arroyo[2]

[1] Centro de Investigación y de Estudios Avanzados,
Instituto Politécnico Nacional,
Mexico

[2] Universidad de Guadalajara,
Centro Universitario de Ciencias Exactas e Ingenieria,
Mexico

matias@cs.cinvestav.mx, salomonarroyo@gmail.com

Abstract. The modeling of cancer metastasis and the immune system (CM-IS) response is of top interest for cancer diagnosis and therapy. CM-IS is a highly complex biological process. From interaction of basic cancer cells emerges the cancer growth and late cancer metastasis. The immune systems reaction for organism protections should avoid the cancer proliferation. The strength of the IS response against cancer spring correlates the success (or not) of the cancer growth. In this paper we outline the use of the Ising model for CM-IS interaction modeling. Ising model is classic in physics, biology and chemistry for the modeling of emergent interaction phenomena. Hence the convenience to use for CM-IS formal and computational intelligence approach.

Keywords: cancer-metastasis, immune-system, interaction, Ising-model.

1 Introduction

CM-IS is a highly complex biological process [8] of top interest for cancer diagnosis and therapy. The next sub-processes are CM-IS involved: the cancer tumor seeding; the cancer cells CC cooperation for tumor growth and metastasis; the immune system cells ISC cooperation strategies for protecting the organism against cancer proliferation; the CM-IS dynamic of fighting –short or long depending on diverse factors. The CM-IS analysis and comprehension requires a multidisciplinary approach to advance regarding the scalability and precision involved in [3]. The last years use of mathematical and algorithmic methods [5, 13, 16] contributes for CM-IS better understanding; as well as for the agile finding of results [17].

Ising model is classic for mathematical modeling of complex interaction phenomena in chemistry, physics and biology [5]. Usually, these phenomena are determined by pressure or temperature as essential thermodynamics parameters.

CM-IS concerns –direct or indirectly– temperature or pressure parameters [13]. But it looks like more complex as an emergent process that involves chemical, physical and biological assorted interaction process [10].

We use Ising model to sketch a formalization of CM-IS. We start with the parametrical description of tissues, and, of some of the effects from CC over tissues for cancer tumor growth or metastasis; as well, the parametrical description of elements of IS process that confront tumor cancer growth and metastasis. We introduce an initial definition of Ising model energy function to formalize the CM-IS interaction: the cooperation among CC or among ISC; as well as the dynamic of CM-IS fighting [3]. On the base of the Ising model formalism computer simulations are practiced and the analysis of results provides conclusions that may help for therapies design.

Next follows a short review of the hallmarks of cancer. Then the literature review on CM modeling, with differential equations, Markov decision processes, agent-based systems and Hamiltonian formulas. The Ising model formalism for CM-IS is introduced and illustrated. A short section for Discussion follows and Conclusions close the paper.

1.1 Hallmarks of Cancer and Conditions to Deploy

Cancer is a multifactorial illness that grows from individual genetic inheritance joins to life style conditions [8, 11]. Genetic difference make different tendencies to cancer deploy [8]. But, cultural factors like habits of life, food quality and person's living conditions, it makes the difference of low or high frequency in a population; or the individual intensity of cancer deploy [6]. Aerobic exercise practice strongly prevents the risk of cancer seed and late deployment [2].

Cancer starts with a disordered replication of malignant cells over a tissue shaping the first solid cancer tumor [8]. The transiton growth factor TGF-β play an important role in the tumor micro-enviroment (TME). At early stages of tumorigenesis can acth like tumor suppressing and tumor promoting later on [8, 4], see Fig. 1. Some cells from the cancer tumor may disseminate on distant tissues making invasions [15]. The CC dissemination is by arterial blood flow or angiogenesis [8]. If the invasion in an organ grows and attracts more CC it becomes successful cancer colonization in the organ. This is cancer metastasis, the fatal step of cancer growth in the live organism [15]. Metastasis implies that cancer is out of a bounded site and is organism spreading. The IS surveillance will have serious difficulties to overcome cancer when metastasis occurs [8]. In Fig. 2 the metastatic process is depicted.

IS reaction to confront and bound cancer spring is clever for health care [5]. The firm IS response makes difficult success of any invasion [3]. Otherwise the weak IS response facilitates the cancer tumor growth and proliferation of metastasis springs [9]. The first IS response is by means of the innate immune system IIS, that recruits macrophages and natural killer cells to eliminate CC invasion [15]. CC antigens are detected by IS cells activating reaction to kill CC. But, mutations of CC complicate the IIS action. A mutant CC is not identified by IIS cells.

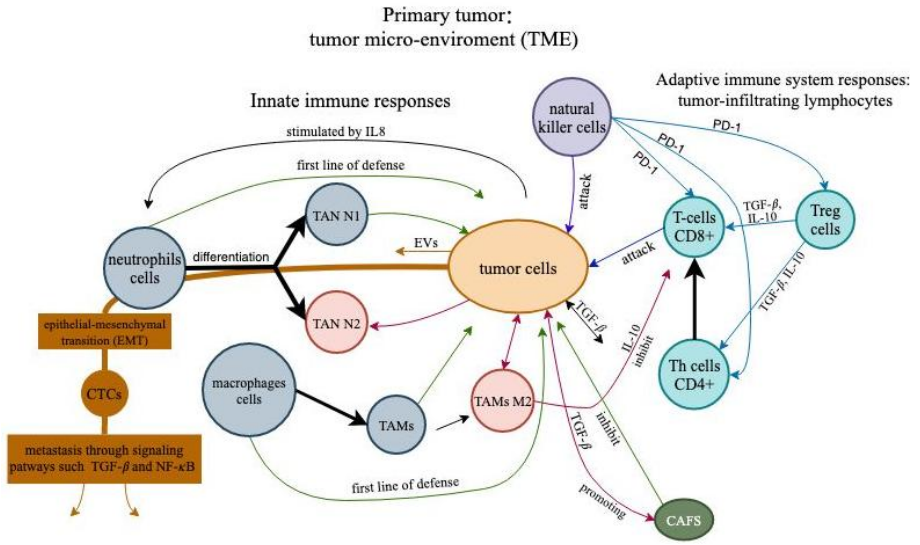


Fig. 1. Interaction of tumor and immune system cells in tumor micro-enviroment (TME).The innate and adaptative immune system interacts in differents forms: neurophils, in form of tumor associate neutrophils (TAN) and macrophages, in form of tumor associate macrophages (TAMs) can be anti-tumoral (TAN N1 and TAMs) or pro-tumoral. Cancer associate fibroblasts (CAFS) initially inhibits the tumor growth, but at the same time can be pro-tumoral trougth TGF-β. The natural killers attacks tumor cells but at the same time regulates the proliferation of adaptative tumors cells.

Therefore, when IIS response is not enough to eliminate cancer tumor, the adaptive immune system AIS action is required. The AIS implements a set of strategies to fight cancer growth and metastasis springs; it activates the recruitment of T cytotoxic cells and other macrophages [7]. A war of biological strategies started. This behavior is available to be formulated then computer simulation on the basis of Ising model.

2 Related Literature Review

In the computational **agent-based** approach each agent is a basic element in a social virtual environment. From the agents interaction usually emerges a behavior not reduced to the linear addition of the agents' behavior. By means of the agent-based model, some CM-IS sub-processes take advantages from computational tools, get scalability and different conditions for testing experiments [7].

In the modeling with **differential equations** [1], cancer growth functions use input parameters that represent the back elements for growth dynamics.

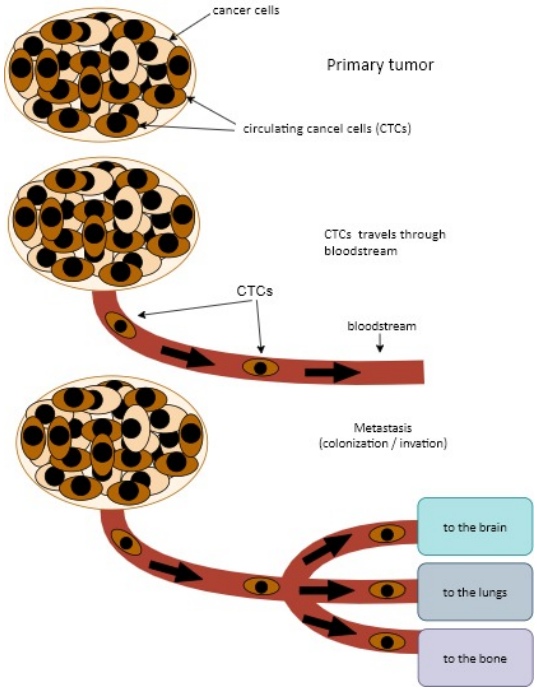


Fig. 2. Diagram of metastatic process. When tumor cells adquire the epithelial-mesenchymal transition, cancer cells can leave the primary tumor in form of circulating tumor cells (CTCs) and travel trough bloodstream. Consequently cancer cells can metastasize (invade) some organs.

As usual, the differential equations depend on the initial and border conditions of the phenomena. However, cancer initial conditions could not be clearly defined, and, border limit conditions less even.

In [14] cancer metastasis from an organ primary tumor is statistically featured. From statistics of 446 patients, the frequency of metastasis for the first site and the sequence of metastasis sites were calculated. As an instance, from breast or prostate cancer the most frequent metastasis site is bone (34%), liver (16%) and lung (15%). Main claim is that the first metastasis strongly influence the deploy sequence of the next metastasis sites; as well, the probability of patient's survival. In addition, a **Markov chain model** of random walks for prediction of the order of spatial metastasis sites is proposed.

3 CM-IS in Systems Biology and Game Theory

The growth of a cancer primary tumor is an illness invasion on live organ tissues. Cancer metastasis is the dispersion of illness, by invasions, to other body organs.

To invade new body tissues (positions), cancer follows kind of strategies. The most usual is the mutation of CC, e. g. PDL1- mutates to PDL1+; it conveys that a mutated CC is no more detected by IS cells that attack it previous mutation. The avoidance of immunity actions accelerates the cancer metastasis. One biology system emerges from primary cancer tumor and late metastasis in a net of lively cancer tumors.

The immune response to a cancer invasion is an IS action of reduction to stop the tumor growth, and the late complete CC elimination. For preventing CC spreading and eventual metastasis, the strengthening of immune surveillance is essential. IIS and AIS construct nets of immunity structures for bounding cancer tumor and metastasis [18] That construction of immune nets obeys kind of strategic composition of IS reductions. The strategies for cancer reductions it involve the recruitment of cytokines T cells and macrophages. Coordination of IS elements is required for the success of surveillance actions. This coordinate immune response is a biology system too.

CM-IS is a two biology system battle to occupy live being tissues. One biology system is the cancer primary tumor and late net of metastasis; the other is built from the immune actions and reactions to preserve health. Hence, CM-IS comprehension requires a system biology approach [10]. As well, the kind of competition to occupy the organs tissues puts CM-IS in a game theory perspective [12]: The goal of each MC-IS gamer is keep control of the organism, by cancer or by health.

4 CM-IS Ising Model Formalism

The Hamiltonian of Ising model [3] for pherromagnetism of spins interaction follows. Value of w_{ij} is the energy interchange between x_i x_j; $-h_i$ is the energy of the field that affects x_i:

$$H = -\frac{1}{2}\sum_{i,j}^{n} w_{ij}x_i x_j - \sum_{i}^{n} h_i|x_i|. \tag{1}$$

The Hamiltonian of Ising model is expressive enough for including the biochemical and biophysical parameters of CM-IS fighting interaction. Notation of parameters in CM-IS process are in Table 1.

Equation (2) is the formal description of molecules in CM-IS; CC molecules is $c_i = -1$ and the IS molecules is $c_i = 1$:

$$x_i = c_i(n_i + r_0^{k_i}). \tag{2}$$

Parameter n_i sets the amount of CC or IS in the molecule. Parameter r_0 sets the rate of TGF-β; and k_i the growth rate of a cancer molecule. The value w_{ij} is the energy interchange between molecules x_i x_j. w_{ij} is calculated from the interaction energy of molecules that cooperate with or confront to x_i x_j:

$$w_{ij} = \sum_{k}^{m} r_{t,x_t^{ij}}. \tag{3}$$

Table 1. Parameters in CM-IS process.

CC/metastasis tumor	Parameter	Immune system	Parameter
Initial tumor cells	n_i	Innate immune system	m_i
TGF-β	r_0	Natural killers	c_i
IL-8	λ	Adaptative immune system	t_i
Tumor derived factors	τ	Fibroblasts	f_i
PD-11	π	Fibroblasts	f_i

Parameter r_t is the coefficient for weighting the strength of tactics of cancer or immune response. The tactics of invasion, reductions, nets and connections among molecules of the CM-IS process. In Table 2 the list of draft percentages for tactics is propose. The higher the percent values of tactics it is the higher the strength of each biology system.

Table 2. List of draft percentages for tactics is propose.

Tactic	Notation	Percentage of strength of cancer / Immune system		
		Weak	Medium	Strong
Invation	r_{in}	0.1	0.4	0.7
Reduction	r_{rd}	0.1	0.4	0.8
Net	r_{nt}	0.2	0.5	0.7
Conection	r_{ct}	0.2	0.5	0.7

Percentages for tactics are just draft values. Current simulations are made as proof of concept with such values. Thoroughly test and adjustments would allow the tune convergence to real values. Patients' real data are required to estimate probabilities on the basis of them. The best the data samples the best the calculi of suitable probabilities; in addition to a wide enough diversity of data is required.

A house-made Netlogo application is used for simulations. Breast cancer primary tumor and bone metastasis simulations are practiced in rough manner. Results show an expected tendency: the equal force of each biology systems splits the percentage of success. And, the proportional strength of one biology system with respect to the other it corresponds to the each other proportional success.

5 Discussion

CM-IS concerns the analysis of biochemical and physiological process for understanding the relationship of an organ primary tumor and the late metastasis

on far organs' tissues. Statistics from data base of patients are being emerging in recent years. CM-IS statistics needs of several years periods to get relevant observations on the cancer evolution, for both metastasis behavior and survival periods. Depending of the type of primary tumor and first metastasis site the evolution is observed with differences. We think that may obey a probability distribution function. It should need a deep analysis for concluding this hypothesis.

The CM-IS biological system can be seen as a game competition to occupy the body. Go game concerns a fight on territory control. First automation that beat top human experts was AlphaGo [16]. Go gaming comparison with CM-IS and formalization by Ising model [3] is an antecedent of current work. Chemo-, radio-therapies or vaccines for cancer treatment may add to CM-IS game theory perspective.

6 Conclusions

From CC interaction emerges the cancer growth and metastasis. The strength of the immune systems reaction against cancer spring and metastasis it determines the success to beat them. CM-IS is a highly complex biological process. We use the Ising model for CM-IS modeling. The CM-IS emergent biological phenomena is well formalized by using this stochastic mathematical tool. A Netlogo program supports the initial simulations. Draft case study involves breast cancer and bone metastasis as the first metastasis site.

Acknowledgment. The work was supported by CONACyT project A1-S-20037. We thank pathologist Mariana P. Arroyo Duarte from Mexican Institute of Social Save IMSS, for her smart advice on CM-IS medical aspects. Misunderstands is all of author's responsibility.

References

1. Altrock, P.M., Liu, L.L., Michor, F.: The mathematics of cancer: integrating quantitative models. Nature Reviews Cancer 15(12), 730–745 (Dec 2015), https://www.nature.com/articles/nrc4029
2. Ashcraft, K.A., Peace, R.M., Betof, A.S., Dewhirst, M.W., Jones, L.W.: Efficacy and Mechanisms of Aerobic Exercise on Cancer Initiation, Progression, and Metastasis: A Critical Systematic Review of In Vivo Preclinical Data. Cancer Research 76(14), 4032–4050 (Jul 2016), https://cancerres.aacrjournals.org/content/76/14/4032
3. Barradas-Bautista, D., Alvarado-Mentado, M., Agostino, M., Cocho, G.: Cancer growth and metastasis as a metaphor of Go gaming: An Ising model approach. PLOS ONE 13(5), e0195654 (May 2018), http://dx.plos.org/10.1371/journal.pone.0195654
4. Barriga, V., Kuol, N., Nurgali, K., Apostolopoulos, V.: The Complex Interaction between the Tumor Micro-Environment and Immune Checkpoints in Breast Cancer. Cancers 11(8), 1205 (Aug 2019), https://www.mdpi.com/2072-6694/11/8/1205

5. Cleveland, C., Liao, D., Austin, R.: Physics of cancer propagation: A game theory perspective. AIP Advances 2(1), 011202 (Mar 2012), https://aip.scitation.org/doi/10.1063/1.3699043
6. Eslami, M., Yousefi, B., Kokhaei, P., Hemati, M., Nejad, Z.R., Arabkari, V., Namdar, A.: Importance of probiotics in the prevention and treatment of colorectal cancer. Journal of Cellular Physiology 234(10), 17127–17143 (2019), https://onlinelibrary.wiley.com/doi/abs/10.1002/jcp.28473
7. Gong, C., Milberg, O., Wang, B., Vicini, P., Narwal, R., Roskos, L., Popel, A.S.: A computational multiscale agent-based model for simulating spatio-temporal tumour immune response to PD1 and PDL1 inhibition. Journal of The Royal Society Interface 14(134), 20170320 (Sep 2017), https://royalsocietypublishing.org/doi/full/10.1098/rsif.2017.0320
8. Hanahan, D., Weinberg, R.: Hallmarks of Cancer: The Next Generation. Cell 144(5), 646–674 (Mar 2011), https://linkinghub.elsevier.com/retrieve/pii/S0092867411001279
9. Jinnah, A., Zacks, B., Gwam, C., Kerr, B.: Emerging and Established Models of Bone Metastasis. Cancers 10(6), 176 (Jun 2018), http://www.mdpi.com/2072-6694/10/6/176
10. Lander, A.: Pattern, Growth, and Control. Cell 144(6), 955–969 (Mar 2011), http://www.sciencedirect.com/science/article/pii/S0092867411002467
11. Lowe, S.S., Danielson, B., Beaumont, C., Watanabe, S.M., Baracos, V.E., Courneya, K.S.: Correlates of objectively measured sedentary behavior in cancer patients with brain metastases: an application of the theory of planned behavior. Psycho-Oncology 24(7), 757–762 (2015), https://onlinelibrary.wiley.com/doi/abs/10.1002/pon.3641
12. Nash, J.: Non-Cooperative Games. Annals of Mathematics 54(2), 286–295 (1951), https://www.jstor.org/stable/1969529
13. Newton, P.K., Mason, J., Hurt, B., Bethel, K., Bazhenova, L., Nieva, J., Kuhn, P.: Entropy, complexity and Markov diagrams for random walk cancer models. Scientific Reports 4(1), 7558 (May 2015), http://www.nature.com/articles/srep07558
14. Newton, P.K., Mason, J., Venkatappa, N., Jochelson, M.S., Hurt, B., Nieva, J., Comen, E., Norton, L., Kuhn, P.: Spatiotemporal progression of metastatic breast cancer: a Markov chain model highlighting the role of early metastatic sites. npj Breast Cancer 1(1), 15018 (Nov 2015), http://www.nature.com/articles/npjbcancer201518
15. Peinado, H., Zhang, H., Matei, I.R., Costa-Silva, B., Hoshino, A., Rodrigues, G., Psaila, B., Kaplan, R.N., Bromberg, J.F., Kang, Y., Bissell, M.J., Cox, T.R., Giaccia, A.J., Erler, J.T., Hiratsuka, S., Ghajar, C.M., Lyden, D.: Pre-metastatic niches: organ-specific homes for metastases. Nature Reviews Cancer 17(5), 302–317 (May 2017), http://www.nature.com/articles/nrc.2017.6
16. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of Go without human knowledge. Nature 550(7676), 354–359 (Oct 2017)
17. Szeto, G.L., Finley, S.D.: Integrative Approaches to Cancer Immunotherapy. Trends in Cancer 5(7), 400–410 (Jul 2019), http://www.sciencedirect.com/science/article/pii/S2405803319301025
18. Vinay, D.S., Ryan, E.P., Pawelec, G., Talib, W.H., Stagg, J., Elkord, E., Lichtor, T., Decker, W.K., Whelan, R.L., Kumara, H.S., Signori, E., Honoki, K., Georgakilas, A.G., Amin, A., Helferich, W.G., Boosani, C.S., Guha, G., Ciriolo, M.R., Chen, S., Mohammed, S.I., Azmi, A.S., Keith, W.N., Bilsland, A.,

Bhakta, D., Halicka, D., Fujii, H., Aquilano, K., Ashraf, S.S., Nowsheen, S., Yang, X., Choi, B.K., Kwon, B.S.: Immune evasion in cancer: Mechanistic basis and therapeutic strategies. Seminars in Cancer Biology 35, S185–S198 (Dec 2015), https://linkinghub.elsevier.com/retrieve/pii/S1044579X1500019X

ISSN 1870-4069

Automatic Cropping of Retinal Fundus Photographs using Convolutional Neural Networks

Gaspar González Briceño[1], Abraham Sánchez[2], E. Ulises Moya Sánchez[2,4], Susana Ortega Cisneros[1], Germán Pinedo[1], Mario S. García Contreras[3], Beatriz Alvarado Castillo[3]

[1] CINVESTAV,
Department of Electrical Engineering,
Mexico

[2] Jalisco Government,
Department of Artificial Intelligence,
Mexico

[3] Centro Médico Nacional de Occidente IMSS,
Department of Ophtalmology,
Mexico

[4] Universidad Autónoma de Guadalajara,
Maestría en Ciencias Computacionales,
Mexico

ggonzalezb@cinvestav.mx,{sortega,gapinedo}@gdl.cinvestav.mx,
{abraham.sanchez,eduardo.moya}@jalisco.gob.mx, drmariosgc@yahoo.com,
eduardo.moya@edu.uag.mx

Abstract. Deep learning (DL) is used in widespread applications including the medical sector. Particularly, Convolutional Neural Networks (CNNs) are architectures commonly used to classify or predicts retinal fundus photographs (RFP). However, the image noise located in the background of the RFP and space location of the retinal Region of Interest (ROI) can affect the performance of these models. One solution to this problem is to make a segmentation of the RFP. In this work we present a Segmentation Model based on CNN (SMCNN) for the cropping process, this model reaches high accuracy levels for test images up to 98%.

Keywords: convolutional neural networks, retinal fundus photographs, segmentation.

1 Introduction

Deep Learning (DL) models are generated by multiple Artificial Neural Networks Layers that can learn the representations of data on multiple levels of abstraction

[6]. Convolutional Neural Networks (CNNs) have proven to be powerful tools for a wide range of tasks, such as image recognition, audio classification, object detection, medical applications, and others [4, 14, 3].

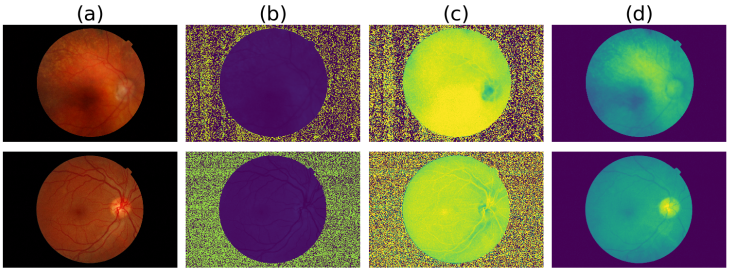


Fig. 1. Background noise representation in two RFPs with apparently good quality. (a) At the top and bottom two RFPs in RGB, (b,c,d) is the representation in HSV of each image respectively. High noise levels can be observed in background once (b) Hue channel and (c) Saturation channel are separated. On the other hand, the (d) Value channel keeps the background values close to 0 (low noise).

Medical applications of CNNs includes clinical and research solutions in the field of eye care, like optical disk segmentation [12], detection of retinal anomalies [7], classification for referable diabetic retinopathy [9], vessel segmentation [13], among others.

Retinal Fundus Photographs (RFPs) are images widely used by physicians to diagnosis an ocular disease. Deficiency in the quality of these images could commit the success of the diagnosis but also the treatment. [1, 8]. In addition, the performance of the DL models could be affected by the difference in the spatial location of the retinal region of interest (ROI) as well as the noise situated in the background, which is not visible to the human eye [15, 8]. These intrinsic problems can be analyzed throw a Hue, Saturation, and Value (HSV) color representation [16], as shown in figure 1(b,c).

In this context, a Segmentation Model based on deep CNNs (SMCNN) is proposed, which predicts and builds the retinal ROI mask of a RFP. This strategy facilitates the retinal cropping process, which helps to avoid the noise of the background and to homogenize the ROI location of each image, all this increasing the image quality used for the CNN models.

The remainder of the paper is structured as follows. The experimental setup, mask generator, SMCNN model, and cropping are described in section 2. Results, can be found in section 3. The discussions are presented in section 4. Finally, conclusions are presented in section 5.

2 Data and Methods

2.1 Classical Mask Generator

In order to build the RFPs' masks used to train, validation and test the SMCNN, we use the method is described in Figure 2. First, the RGB images are turned into the grayscale in order to avoid the background noise of the RFPs.

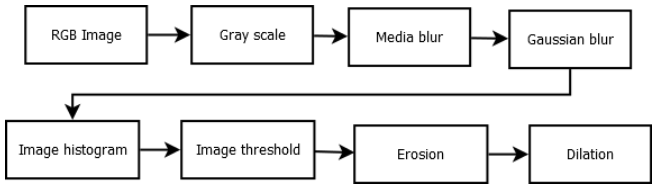


Fig. 2. Flowchart of the classical mask generator for train, validation and test.

Second, medium blur and high-value Gaussian blur filters are applied with the goal of reducing the variance between the pixel values, this produces a uniform distribution throughout the image. Then, the contrast histogram is generated, which works to know the threshold between the minimum and maximum values, identifying thus the upper and lower limits for the image binarization. Finally, erosion and dilatation filters are added to smooth the edges of the retinal ROI. As a result, we obtain some preliminary masks and these ROI masks were selected and corrected by hand.

2.2 Mask Dataset

A total of 500 binary masks were generated with the classical mask generator(aided by human) from the Kaggle Diabetic Retinopathy Detection dataset [2]. The 500 images were used to train and validate the SMCNN, which predicts the masks of new RFPs. The quantity of images used is shown in Table 1. Binary masks obtained from the classical mask generator are resized with a 1024x1024 resolution, and then they are correlated with their grayscale images counterpart, which in turn proceeds from the RFPs in RGB.

Table 1. Dataset used for the SMCNN, based on Kaggle dataset.

Dataset	No. of Images	Train	Validation	Test
Kaggle	500	280	120	100

2.3 SMCNN Training

In Figure 3, the training an validation process are presented. The configuration of the U-Net architecture [10] is based on a back-propagation algorithm[11] and Adam optimizer [5]. We use a GCloud instance with a V-100 GPU. Then, a cropping process was generating to obtain the same spacial proportion, see figure 4 as an example.

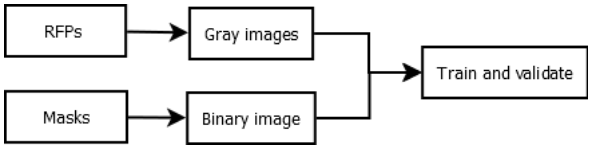


Fig. 3. SMCNN training flowchart process.

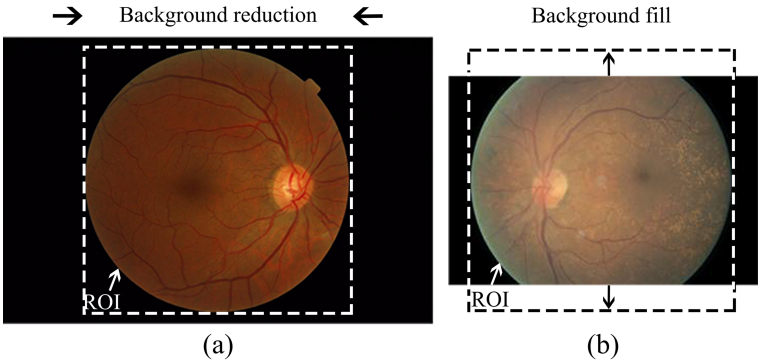


Fig. 4. Cropping process, (a) background reduction, maintain ROI's aspect ratio, (b) background reduction and fill of ROI.

3 Results

3.1 Dataset Creation With Classical Mask Generator

Examples of the dataset are shown in Figure 5. Note that, only the results with the best approximation of the shape of the retina were selected manually.

3.2 SMCNN Performance

The performance of the SMCNN model made to predict the masks of RFPs is presented at table 2. High levels obtained in the accuracy for training, validation,

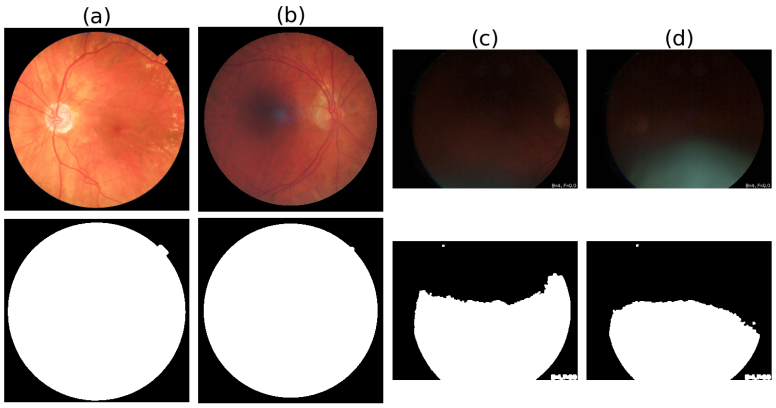


Fig. 5. Example of the results of our personalized technique for the creation of masks. Figures (a, b) are some of the successful results therefore they were selected for the training, validation and testing process. Meanwhile figures (c, d) are unsatisfactory so they were discarded.

Table 2. Train, validation and testing results.

Train		**Validation**		**Test**	
Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
0.99	0.0030	0.98	0.0031	0.98	0.01898

and test are interesting results, due that they represent the convergence rate between the original RFP and the one predicted by our SMCNN model.

Figure 6 shows an example of a result of the cropping process with SMCNN, using low contrast images, in addition, is compared with a classical mask generator.

4 Discussion

Techniques described in this work represent a solution to get the retinal ROI of diversity images obtained with different methodologies. Nonetheless, a mask obtained by this methodology could be incomplete if the image quality is over or under certain quality values. The quantification and demonstration of these quality levels are proposed as future work.

5 Conclusions

In this work, a retinal segmentation model based on CNN was introduced. ROI mask obtained of this model facilitates the cropping process, avoiding the noise located in the background of the RFPs and homogenizing the location of the

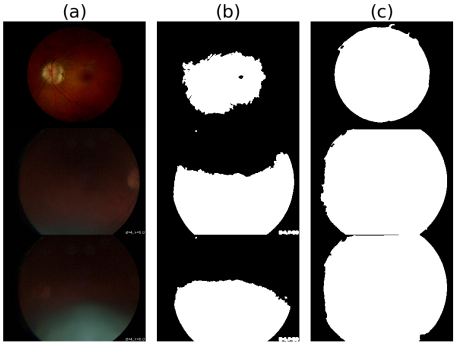


Fig. 6. Example of the comparison between a classical mask generator and DL segmentation process. (a) RFP in RGB. (b) Result using classical technique like Figure 5. (c) Result of segmentation process based on DL.

retinal ROI in the image space. Results show an accuracy of up to 98% on the test set. We hope to use transfer learning on another dataset in order to generate more segmented samples form other datasets.

References

1. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B.: Medgan: Medical image translation using gans. Computerized Medical Imaging and Graphics 79, 101684 (2020)
2. EyePACS: Diabetic retinopathy detection competition. https://www.kaggle.com/c/diabetic-retinopathy-detection/ (2015)
3. Greenspan, H., Van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. vol. 35, pp. 1153–1159 (2016)
4. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. vol. 35, pp. 221–231 (2013)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR 1412.6980 (2014)
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436 (2015)
7. Li, Q., Fan, S., Chen, C.: An intelligent segmentation and diagnosis method for diabetic retinopathy based on improved u-net network. Journal of Medical Systems 43(9), 304 (2019)
8. Moya-Sánchez, E.U., Sánchez, A., Zapata, M., Moreno, J., Garcia-Gasulla, D., Parrés, F., Ayguadé, E., Labarta, J., Cortés, U.: Data augmentation for deep learning of non-mydriatic screening retinal fundus images. In: International Conference on Supercomputing in Mexico. pp. 188–199. Springer (2018)
9. Pires, R., Avila, S., Wainer, J., Valle, E., Abramoff, M.D., Rocha, A.: A data-driven approach to referable diabetic retinopathy detection. Artificial Intelligence in Medicine 96, 93–106 (2019)
10. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer (2015)

11. Rumelhart, D.E., Durbin, R., Golden, R., Chauvin, Y.: Backpropagation: The Basic Theory, p. 1–34. L. Erlbaum Associates Inc., USA (1995)
12. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. Pattern Recognition and Image Analysis 27(3), 618–624 (2017)
13. Son, J., Park, S.J., Jung, K.H.: Retinal vessel segmentation in fundoscopic images with generative adversarial networks (2017)
14. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 25(1), 44 (2019)
15. Wang, J., Bai, Y., Xia, B.: Feasibility of diagnosing both severity and features of diabetic retinopathy in fundus photography. IEEE Access 7, 102589–102597 (2019)
16. Wu, Y., Zheng, J., Song, W., Liu, F.: Low light image enhancement based on non-uniform illumination prior model. IET Image Processing 13(13), 2448–2456 (2019)

ISSN 1870-4069

Random Forest and Deep Learning Performance on the Malaria DREAM Sub Challenge One

Didier Barradas-Bautista

King Abdullah University of Science and Technology, Catalysis Center, Saudi Arabia

didier.barradasbautista@kaust.edu.sa

Abstract. In several countries, vector borne diseases play a significant role in the burden of public and individual health. Mosquitoes transmit diseases such as dengue and malaria. Malaria is a disease caused by *Plasmodium* parasites and transmitted by *Anopheline* mosquitoes species. In 2017, WHO estimates 200 million cases, affecting mainly children under five years old. On 2018 in Mexico, reported 799 Malaria cases showing a more similar trend for 2019 so far. Among the effort to eradicate malaria, a crowd-sourced event called DREAM challenge, where participants have to produce machine learning models and strategies on biological data. Also, the baseline truth is not released during the challenge. On this DREAM challenge, the participants can use any approach any design to build a predictor to predict the clearance of and concentration of artemisin (the standard antimalarial drug). On this work, I discuss differences in my strategy over sub challenge one in which my predictions ranked among the top three performers.

Keywords: DREAM challenge, ensemble methods, malaria.

1 Introduction

1.1 Malaria a World Wide Problem Almost Neglected

Malaria of "several diseases transmitted by insects, also known as Vector born diseases, with a considerable impact on human health [13]. It is caused by protozoans from the genus *Plasmodium* and transmitted by *Anopheles* mosquitoes. It has been more than 100 years since Ronald Ross discovered the *Plasmodium* parasites on Anopheles mosquitoes guts and understood the spreading mechanism of this disease [22]. But in Malaria is still an ongoing problem, in 2017, the World Health Organization (WHO) estimates 200 million cases, affecting mainly children under five years old [20]. In 2018 in Mexico reported 799 Malaria cases showing a similar trend for 2019 so far [23]. *P. falciparum* and *P. vivax* are the most common cause of Malaria in Africa and America, respectively.

Over the years, the global number of Malaria cases has dropped thanks to initiatives like "roll back malaria" [9, 18] or "The Malaria day in the Americas" [1].

Plasmodium protozoans have a very complicated life cycle where it changes several times the proteins in its surface to avoid recognition from the mosquito and human immune system [15]. The continue adaptation makes it difficult for the development of a highly effective vaccine or drug [16]. Moreover, the problem can become more complicated with the rise of resistant mosquitoes to insecticides and *Plasmodium* parasites resistant to drugs [25].

As part of the effort to fight Malaria, crowd-sourcing projects can help to develop new antimalarial drugs or to predict their activity [19]. Also, crowd-sourced challenges can help understand the mechanism of the underlying biology of Malaria, like the Malaria DREAM Challenge [11, 7].

1.2 Palliative Treatment of Malaria

The WHO Patients diagnosed with Malaria are preferably treated with artemisinin combination therapies [2]. Artemisinin is a drug that eliminates the majority of *Plasmodium* parasites, while the remaining parasites are eliminated by other partner drugs [8]. On the other hand, drug-resistant *Plasmodium* is a recurrent problem [5].

1.3 The Malaria DREAM Sub Challenge One Design

With the advent of the so-called OMIC techniques, a high throughput machine can process hundreds or thousands of samples producing big data over complex biological scenarios. In this situation, computational models and methods are the best tools to understand the underlying biological process on these. The DREAM challenge is an open science effort inspired by the crowd-sourced creation of such computational tools and analysis from the scientific community.

The general objective of the challenge was to predict the artemisinin resistance level. Thus the DREAM Malaria challenge was divided into two parts. The first challenge consisted of the concentration values for half of the inhibitory concentration of artemisinin (IC50). The second part was dedicated to predict the patient status of clearance from parasites post-treatment. In both cases, the data consisted of expression changes of all the genes of the *Plasmodium* parasite. Still, the first part used *in-vitro* to predict *in-vitro* expression; on the second part, the *in-vitro* data was used to predict *in vivo* response.

2 Methods

The organizer supplied the data, that consisted of expression data form the 5542 genes for the *Plasmodium* parasite and four additional descriptors. These four descriptors provided information about the time of recollection ("Timepoint"), type of treatment ("Treatment," describing the use of artemisinin), bio-repeats ("Biorep"), and the id of origin.

To preprocess the data, I used the pandas library [17]. I used two different approaches to build the models, first I used sci-kit learn [21, 24] for the random

forest model and hyperparametric search and second, I used Tensorflow [3] and Keras [6] to build the neural network model.

2.1 Data Cleaning

The data for this sub-challenge was complete since it does not have any empty values. I converted the categorical columns "Treatment," and "Timepoint" to dummy variables to manage this relevant information. I dropped the isolate, and Biorep features the training and testing data set. Isolate feature is an identification label that not informative for training since a different set of isolates constitute the testing data. The Biorep feature is another id label that is not included on the independent testing set; thus is not informative. The training data was used to calculate the mean and standard deviation to use for later scaling of itself and the testing data. I trained the models from this standardized data.

3 Result

On the DREAM challenges, participants should provide the source code for the final models used for predictions. This setup is a two-fold advantage for all the community since the idea is to share the ideas and secure the reproducibility of the code. Thus my code is available on the challenge page[1].

3.1 Random Forest Model

To obtain the best random forest model, I performed a random search of hyperparameters with five-fold cross-validation. A random search of parameters allows testing several combinations of decision trees with a minimal number of samples and splits. According to the search the best parameters for a random forest models are 6333 trees taking all the data to build each tree, with a maximum number of depth features set to 30, a minimum of ten samples to node split and a minimum of 8 samples to be considered a leaf. The produced model achieved 69.59% accuracy with 0.3899 degrees of Mean average error (MAE), and 0.2746 degrees of mean squared error (MSE), see Fig. 1A. The detailed script of the search is available on GitHub[2].

3.2 Neural Network Model

A fully connected neural network, with five layers with 640 neurons to deal with the several combinations of input, four layers with 64 neurons to condense the previous information and a final output layer.

I used the Rectified Linear Unit as an activation function and RMSprop with the default learning rate as the optimizer. This architecture had 5,243,329

[1] https://www.synapse.org/#!Synapse:syn20609261
[2] https://github.com/D-Barradas/random_Hyperparameter_Search-

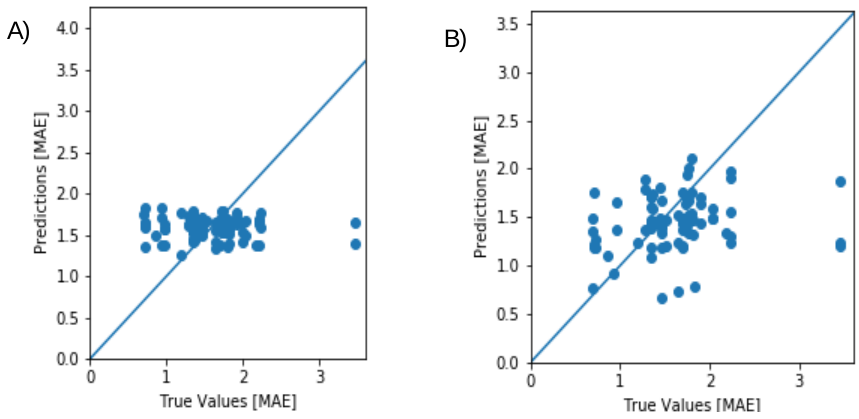


Fig. 1. Scatter plot of predicted vs true values. A) Resulting prediction distribution from the random forest algorithm. B) Resulting prediction distribution from the deep neural network.

of trainable parameters. The batch size set to 2 samples. I trained the model on the scaled data, and the model stopped around 20 epochs because there was an early stop clause for when the learning rate did not improve. The resulting model had an accuracy of 72.20% but with bigger MAE and MSE, 0.4143 and 0.349 respectively (Fig. 1B).

4 Discussion

The typical biological data set contains measurements of individual genes of organisms; in the case of the DREAM malaria Challenge, the data is the whole expressed genome of the *Plasmodium* parasite. This means the data reflected the actual genetic response of the parasite in the presence of a drug. Hence, I considered this problem as a multivariate data set. The underlying problem is, in fact, an interaction network, and searching for correlations among the 5542 genes is computationally a challenge. Understanding that is a network of interactions discards some approaches to predict the actual response, such as lineal regression or support vector machines. To the best of my knowledge, ensemble methods such as Decision trees and densely connected neural networks are the best algorithms to build predictions from complex feature relationships.

On that line, I searched for the best random forest predictive model. The resulting model had roughly one tree per gene where at least eight genes are needed to set the split for a new leaf of the decision tree, with a maximum depth of 30 related genes. This kind of arrangement of the decision trees is a consequence of the underlying network previously mentioned.

Thus, we could say that up to a combination of 30 genes are involved in some way to the response to artemisinin. Moreover, overlooking the feature

Table 1. Top 10 features ranked according the importance from the random forest model.

Description	Genes ID	Importance
Unknown function	PF3D7_1326200	0.004056
Unknown function	PF3D7_1217000	0.003661
Unknown function	PF3D7_0108200	0.003535
Sec1 family protein	PF3D7_1034000	0.003250
Unknown function	PF3D7_1307700	0.003152
O-fucosyltransferase 2	PF3D7_0909200	0.002717
Unknown function	PF3D7_0626200	0.002685
MORN repeat protein	PF3D7_1426400	0.002528
Unknown function	PF3D7_0502500	0.002503
Liver merozoite formation protein	PF3D7_0602300	0.002480

importance (Table 1), we can obtain a list of the genes and their functions. It is known that about half of the *Plasmodium falciparum* genome encodes conserved proteins of unknown function [4], so the most of the features used have a biological unknown function. However on Table 1 a few features have a function assigned that are related to membrane dynamics on a particular moment in the parasite life cicle. Sec1 proteins plays a role on membrane exocitosis [12], MORN repeat proteins are associated with cell budding and nuclear division [10], and O-fucosyltransferase 2 is required to the proper assembly of trafficking of proteins [14].

All these functions together tied together are indicators of the schizogony stage (asexual reproduction by fission) which is previous to the release of merozoites hence among the features we can find the "Liver merozoite formation protein" among the top ten features used on the random forest model. These genes could become new markers or drug targets.

Spite this apparent predictive power, the predictions with ensemble methods become constrained and somewhat limited by the number of genes that can be combined on each of the decision trees. A highly connected neural network could, in theory, overcome the problem. I trained a deep neural network with a funnel-like architecture, diminishing the number of neurons towards the output layer. The resulting architecture also had batches of two samples, generating a framework that searches for the best correlation on binary interactions.

Using deep learning comes at the cost of losing the precise descriptive power of decision trees. But on the other hand, we can monitor the training of the model through the use of quality metrics like MAE and MSE. The curves for training and validation showed a very early stoppage (before 20 epochs). For MAE (Fig. 2A) as for MSE (Fig. 2B), the validation error was slightly above the training error, indicating the possibility of over-fitting. Probably a similar training occurred for the best random forest model.

The random forest used the whole data set to build each tree, while the deep neural network was driven by all the combinations of the entire data-set on

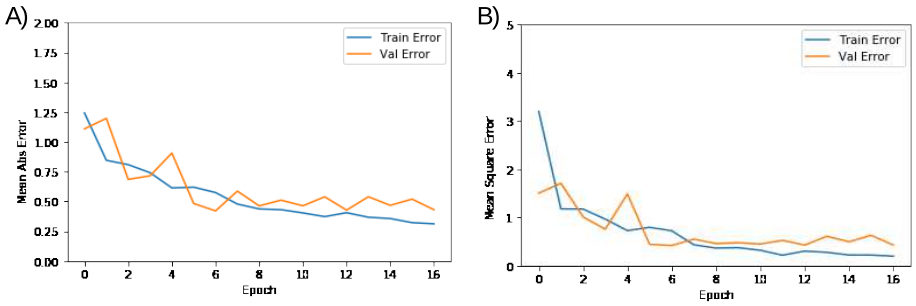


Fig. 2. History of the neural network training on the data. A)MAE values during the neural network training. B)MSE values during the neural network training.

the first layers. However, the prediction from the neural network has a broader range, reflected in the difference of MSE with the random forest. This wider range is also the reason for the 3.39% increase on the accuracy compared to the random forest.

5 Conclusion

Multivariate data requires an algorithm with complex decision-making schemes. The kind of model that can present this is ensemble methods and neural networks. I based my choice to use deep neural networks over the random forest model on the accuracy and the visual inspection of the prediction range. While using random forest explains the decision making, it requires thousands of trees for the prediction. Then deep learning is an alternative option that can fit multivariate data with the clear drawback of becoming unable to know the detail of the combinations made on each neuron. Overall the deep learning model made predictions close enough to be the third place on the sub-challenge. Further feature engineering and selection could improve both models. Reviewing the feature importance, we can discover new drug targets or assign functions that remained unknown for some genes so far.

Acknowledgment. The work is supported by KAUST Catalysis Center. I want to thank the KAUST Supercomputing Laboratory (KSL) for allowing me to use the resources available, especially the Shaheen and Ibex supercomputers.

References

1. Campeones de la malaria día del paludismo en las Américas

2. Guidelines for the treatment of malaria. Geneva (2006)
3. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., a nd Sherry Moore, R.M., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015)
4. Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O.S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J.C., Lawrence, C., Li, W., Pinney, D.F., Pulman, J.A., Roos, D.S., Shanmugasundram, A., Silva-Franco, F., Steinbiss, S., Stoeckert, C.J., Spruill, D., Wang, H., Warrenfeltz, S., Zheng, J.: EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Research 45(D1), D581–D591 (2017)
5. Barnes, K.I., White, N.J.: Population biology and antimalarial resistance: The transmission of antimalarial drug resistance in Plasmodium falciparum, vol. 94 (2005)
6. Chollet, F., others: Keras (2015), https://keras.io
7. Davis, S., Button-Simons, K., Bensellak, T., Ahsen, E.M., Checkley, L., Foster, G.J., Su, X., Moussa, A., Mapiye, D., Khoo, S.K., Nosten, F., Anderson, T.J.C., Vendrely, K., Bletz, J., Yu, T., Panji, S., Ghouila, A., Mulder, N., Norman, T., Kern, S., Meyer, P., Stolovitzky, G., Ferdig, M.T., Siwo, G.H.: Leveraging crowdsourcing to accelerate global health solutions. Nat. Biotechnol. 37(8), 848–850 (2019)
8. Dondorp, A.M., Nosten, F., Yi, P., Das, D., Phyo, A.P., Tarning, J., Lwin, K.M., Ariey, F., Hanpithakpong, W., Lee, S.J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S.S., Yeung, S., Singhasivanon, P., Day, N.P.J., Lindegardh, N., Socheat, D., White, N.J.: Artemisinin resistance in Plasmodium falciparum malaria. N. Engl. J. Med. 361(5), 455–467 (2009)
9. Feachem, S.R.: Roll Back Malaria: an historical footnote. Malaria Journal 17(1), 433 (2018)
10. Ferguson, D.J.P., Sahoo, N., Pinches, R.A., Bumstead, J.M., Tomley, F.M., Gubbels, M.J.: MORN1 Has a Conserved Role in Asexual and Sexual Development across the Apicomplexa. Eukaryotic Cell 7(4), 698–711 (2008)
11. Ghouila, A., Siwo, G.H., Entfellner, J.B.D., Panji, S., Button-Simons, K.A., Davis, S.Z., Fadlelmola, F.M., The DREAM of Malaria Hackathon Participants, Ferdig, M.T., Mulder, N.: Hackathons as a means of accelerating scientific discoveries and knowledge transfer. Genome Research 28(5), 759–765 (2018)
12. Halachmi, N., Lev, Z.: The Sec1 Family: A Novel Family of Proteins Involved in Synaptic Transmission and General Secretion. Journal of Neurochemistry 66(3), 889–897 (2002)
13. Institute of Medicine (US) Forum on Microbial Threats: Vector-Borne Diseases: Understanding the Environmental, Human Health, and Ecological Connections, Workshop Summary. The National Academies Collection: Reports funded by National Institutes of Health, National Academies Press (US), Washington (DC) (2008), http://www.ncbi.nlm.nih.gov/books/NBK52941/
14. Lopaticki, S., Yang, A.S.P., John, A., Scott, N.E., Lingford, J.P., O'Neill, M.T., Erickson, S.M., McKenzie, N.C., Jennison, C., Whitehead, L.W., Douglas, D.N., Kneteman, N.M., Goddard-Borger, E.D., Boddey, J.A.: Protein O-fucosylation

in Plasmodium falciparum ensures efficient infection of mosquito and vertebrate hosts. Nature Communications 8(1), 561 (2017)
15. Matuschewski, K.: Getting infectious: formation and maturation of Plasmodium sporozoites in the Anopheles vector. Cellular Microbiology 8(10), 1547–1556 (2006)
16. Matuschewski, K.: Vaccines against malaria-still a long way to go. The FEBS Journal 284(16), 2560–2568 (2017)
17. McKinney, W.: pandas: a Foundational Python Library for Data Analysis and Statistics
18. Nabarro, D.: Roll back malaria. Parassitologia 41(1-3), 501–504 (1999)
19. OpenWetWare: Opensourcemalaria:faq (2016), https://openwetware.org/
20. Organisation Mondiale de la Sante: World malaria report 2018 (2018)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python, vol. 12 (2011)
22. Rajakumar, K., Weisse, M.: The Centennial Year of Ronald Ross' Epic Discovery of Malaria Transmission: An Essay and Tribute, vol. 43 (1998)
23. Secretaria de Salud: Boletin epidemiologico
24. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., Mueller, A.: Scikit-learn, vol. 19 (2015)
25. White, N.J.: Antimalarial drug resistance. Journal of Clinical Investigation 113(8), 1084–1092 (2004)

Regular Papers

ISSN 1870-4069

COVID-19 Pandemic: An Overview of Machine and Deep Learning Methods for Analysis of Digital Media Texts

Nouf Matar Alzahrani

Albaha University,
College of Computer Science and Information Technology,
Saudi Arabia

noufalzahrani@bu.edu.sa

Abstract. The global efforts to generate information during the COVID-19 outbreak are awe-inspiring. Governments are attempting to make sure people's health is safe and sound during this epidemic by automatically processing the giant amount of online text. This text helps governments to make appropriate policies promptly by understanding public opinion at a suitable time to avoid outrageous consequences. The social media platforms play a significant role in fostering healthy online public communities, particularly for the user to user interaction in such pandemic circumstances. However, manually processing and analyzing a huge amount of data is a troublesome task. So, the study aims to provide a comprehensive analysis of the methods used to automatically process and analyze the digital media text. Moreover, the study also sheds light on the traditional machine learning and deep learning algorithms used to monitor users' activity, attitude, and response to ample amounts of information on various social media platforms during the outbreak of COVID-19. The study concludes that each digital platform has been used for different goals, such as communication and thus user's response is different on each digital media platform.

Keywords: social media, machine learning, deep learning, natural language processing.

1 Introduction

The World Health Organization (WHO) officially announced the COVID-19 as the name of this new disease that poses a severe global threat[1]. A recent study [1] highlighted that worldwide hazards are unified, especially the case of the COVID-19 epidemic exhibits how the prevalence of information can be crucial in such pandemic situations.

The flow of a huge amount of information, in particular, fake news on the internet can significantly augment the epidemic process because it influences people and fragments the social response [2].

[1] WHO: Naming the coronavirus disease (COVID-19) and the virus that causes it.

For example, the American news channel is known as CNN[2] published an article showing the possibility of lock-down in Lombardy (a region in northern Italy). Since the news was published a few hours earlier than the official communication from the Italian Prime Minister, people started to travel from Lombardy to other regions. As a result, all the airports, train stations were overwhelmed before implementing the lockdown, and the government initiative was interrupted. It opens a room for many research questions, but the most important question is to investigate the sources where people receive information and make decisions. A recent study [3] highlighted that when people make a decision after consuming information, such decisions have an impact on their attitude. Finally, the COVID-19 epidemic clearly illustrates how social and traditional media influence people's decisions, in particular, and the complete society. Therefore, governments and other organizations are interested in designing automatic tools to understand public opinion in such catastrophic situations to provide people with the best health services and take proper steps to control the epidemic.

Easy access to the internet and communication technologies have provided people a significant opportunity to connect with each other and communicate regardless of the distance. The mechanism to monitor communication on social media platforms, especially when governments are imposing lock down restrictions due to the COVID-19 epidemic is demanding. The reason is that in such situations, information on social media platforms influences people's behavior and it can have a substantial impact on people's decisions. For example, information without proper evidence can easily disrupt government countermeasures especially when the situation is a severe threat to public health. Therefore, it requires automatic models to predict people's attitude towards epidemics such as COVID-19. These predictive models can be extremely fruitful to measure the people's responses after consuming traditional news steam and social media content [2, 4, 5].

People can have direct access to an ample amount of information on various platforms. These platforms include traditional news channels, multiple social media networks such as Facebook, Twitter, YouTube, Instagram. This direct access to information may contain rumors or unverified news. However, new techniques [6] have been proposed to target the specific audience based on the user's preference and response. Although the algorithms can mediate in finding specific audiences and to promote content, however, this may lead to elicit misleading information phenomena [6]. Thus, rumors or unvaried content can provoke public discussion on social media platforms [7, 8], and impact policy-making process, in particular, when the issues arouse controversy and help to develop social perceptions [9, 10].

Online social media platforms have an active wide range audience that provide the worldviews on various topics to its users [11]. These broader views often neglect the conflicting facts [12, 13] and therefore people establish groups that share the same narratives [14, 15]. As a result, when the degree of polarization is intense, it is more likely to disseminate false information [16, 17]. Previous research [18] showed that the factually incorrect news scatter with a higher velocity than the verified news. Nonetheless, the fact checking phenomena might be limited to a specific social media platform since the term "Fake News" may be misleading due to the face that the unpleasant news for a particular group can be annotated as a factually incorrect content

[2] CNN: Italy prohibits travel and cancels all public events in its northern region to contain coronavirus.

[19]. The study aims to investigate the machine learning and deep learning methods used to process the digital media in the COVID-19 epidemics. A series of recent studies [10, 18, 20] have indicated that each platform has a different mechanism and provides different services to its users. For example, some platforms are used to share pictures, and stories. In contrast, some online platforms are used to keep people up to date by publishing latest news. To keep this in mind, in this research we presented an overview of the methods that are being used to automatically analyze and process the online text in the pandemic of COVID-19.

The remaining sections of the paper are structured as follows: applications of AI in digital media mining are demonstrated in section 2. In section 3, the methods used to process online text are provided in detail. Section 4 contains the analysis of various social media platforms. Furthermore, discussion and questions for further research are discussed in section 5 and 6 respectively. Finally, the paper concludes with the conclusion in Section 7.

2 Applications of AI in Mining Digital Media Texts

Social media platforms have successfully become an important medium of communication. These platforms are extremely helpful for people to communicate with each other, share their emotions in the form of posts and know up to date worldwide information. However, people are generating a giant amount of data each second on these platforms. For example, the World Economic[3] report that by 2020 it is expected to generate 44 zettabytes. So, it implies the number of stars[4] in the whole plant are 40 times less than bytes generated on the internet in 2020. Nonetheless, it requires automatic tools to process such amounts of data.

Covid-19[5] pandemic has restricted people at home and has physical distance to prevent a catastrophic situation. People have been using Social media platforms as a main communication tool and generate a substantial amount of data during the coronavirus epidemic. For example, a recent study [20] shows that, in every minute, there are approximately 3,3 million Facebook posts. In addition to this, the study [20] also reveals that, approximately 448,800 Tweeting posts are published on Twitter every minute.

It is evident to note that the flow of information can be seen on social media platforms within a few seconds after a crisis occurs. Nonetheless, manually processing to get useful insights from this huge amount of data is quite crucial due to a number of reasons. For example, each platform has a different flow and structure of information, such as Facebook Posts, Twitter Tweets, YouTube video, Instagram pictures. Moreover, the text used on these platforms is unstructured, unorganized, and heterogeneous (different languages text, acronyms, and misspellings). The multiple heterogeneous text features introduce a noise in the data that makes text different to process and understand to obtain fruitful insights and the conveyed message.

[3] https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

[4] https://www.visualcapitalist.com/how-much-data-is-generated-each-day/.

[5] https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Table 1. Comparison of the data gendered in 2017 and 2020 in every minute [22].

Year	2017 (in millions)	2020 (in millions)
Facebook new users	360	400
Google searchers	3.8	4.2
Tweets constructed	448,800	480,000
Instagram images uploaded	66,000	60,000
YouTube videos viewed	4.2	4.7
Emails sent	150	200

Therefore, various applications of machine learning algorithms (supervised and unsupervised machine learning) came into being. These algorithms are significant in not only understanding the text and provide meaningful results, but also automate the entire processing.

3 Methods for Automatic Data Processing

In this section, we provide an overview of different methods used to automatically analyze the online text. These methods [50, 60] have been used to understand public sentiment, response and attitude towards COVID-19 [58, 59] epidemic.

3.1 Supervised Learning

In supervised machine learning, annotation (a process to label data) is required which is mostly done by humans. For example, the text on social media platforms about COVID-19 may contain some characteristics in it. So, the label of the text post can be real information or fake information in a binary class scenario. Furthermore, the supervised learning algorithms process the text in the form of features. The feature is defined as an important piece of information that can help algorithms to correctly classify the information according to its labeled class. There are multiple kinds of features, in particular, each task can be solved with unique features since there is no standard set of features that can help the algorithm to provide better results. In Natural Language processing tasks, these features can be factual based, emotional, sentiment based, word embeddings, lexical features. Moreover, some other features such as time, location, the source of information are also taken into account as features. Finally, the main goal of this learning is to train an algorithm so that it can correctly predict the label of the unknown data.

A number of studies have investigated various machine learning algorithms [23, 24, 25]. These algorithms such as Naive Bayes [26, 27], Support Vector Machine (SVM) [25,28], Random Forests [29], and Logistic Regression [30, 31] and Decision trees [32]

have been used in various text classification application [57] especially in COVID-19 related text processing. To process social media especially on Twitter, recent studies used certain tags to analyze the twitter posting [33], in particular, the tag "breaking news" and "news" has been used to differentiate breaking news twitter posting [34]. Similarly, another study classified tweets using the word "cardiovascular disease" as a key feature [35].

In supervised learning, the algorithms are trained on labelled dataset. Substantial human efforts are required to annotate the dataset correctly. However, since automatic tools can be used to label the data, the risk of incorrect annotation is present which might lead an algorithm to make a wrong prediction. Labeling the dataset in crisis such as COVID-19 is a troublesome task [36]. Such situations require instant information to make new policy decisions for the government organizations. Furthermore, making use of a classifier trained on different data (not related to COVID-19) will not predict and provide required insights of the information related to COVID-19 since it did not learn the pattern of the COVID-19 related data. Therefore, this requires addressing the annotated data limitation by suggesting other methods. Finally, unsupervised learning techniques are introduced to meet the data labeling related challenges.

3.2 Unsupervised Learning

Another machine learning method known as unsupervised methods that require unlabeled data for training [37]. These algorithms are powerful to provide valuable information about the data using clustering techniques. Clustering techniques are significant when seekers are ignorant about what information can be obtained from data. This is the particular case when clustering techniques have been used in the COVID-19 text processing to understand the situation globally [38]. It significantly not only reduces the human efforts to label the data, but also provides data insights (by grouping similar data together) promptly that can be helpful to control any outrageous situation such as COVID-19. Similarly, a recent study used spatio-temporal distribution to spots potentially COVID-19 affected areas using unsupervised learning [39, 54].

Another clustering technique known as soft clusters have been used [40] for predicting COVID-19 patients. This method permits items (in this case patients) to join other clusters with variant degrees simultaneously. In addition to this, the approach has been applied [42] in predicting COVID-19 outbreak by finding the tweets similarity using different features like word selection, length of the tweets, time and location.

3.3 Deep Learning

Artificial intelligence has substantially advanced the techniques to perform a number of tasks automatically [53]. The deep learning (DL) methods [51, 53, 55] are based on human brain structure. The emergence of DL has provided a nudge to the whole research community in solving challenging tasks that require huge computation power and resources.

The deep learning methods are based on various hidden layers (deep layers) and each layer contains a number of neurons. Each neuron in the deep layer is activated and passes the information to the next neuron. This process continues until the final is

provided. These deep layers are extremely helpful in identifying the hidden complex structures in large datasets.

Another important aspect of deep learning method is extremely good in finding the suitable features automatically. These methods are given raw data as an input according to the required data type. For example, in Natural language processing, word embeddings of the data are given as an input to the deep learning models instead of raw data. In contrast, traditional machine learning techniques, we need to perform careful feature analysis on raw data to investigate what features contain more information and are valuable for the algorithm to correctly make the prediction.

There are several deep learning methods such as Convolutional neural networks, recurrent neural networks, which have been used in predicting medical symptoms [52, 56], particularly, the COVID-19 outbreak [42,43]. Deep convolutional nets performed extremely well in processing images related data, however, LSTM which is another type of recurrent nets have been used in classifying the sentiments related to COVID-19 [44] in online discussions. The method [44] used sequential data analysis on text to identify public sentiment on coronavirus diseases. Another study [45] used various neural networks to highlight COVID-19 patients who had pneumonia on their chest. A recent study [46] is exclusively focused on analyzing digital media text on social media platforms, checking the facts of the information.

Deep learning methods require a substantial amount of data for training. Covid-19 generated a substantial amount of data related to different diseases, death and the public opinion. For example, a study [47] used BERT technique to process Twitter posting using unlabeled data. Furthermore, similar studies [48, 49] used deep methods in the authorship attribution task. Finally, it is important to mention that deep learning methods have a variety of other applications to recognize hidden structure using images, audios and video related data and provide extremely good results to automatically perform multiple tasks.

4 Analysis of Digital Media Texts

Social media platforms have become an important part of our society because these platforms provide technologies to communicate with each other, engage with worldwide information [2, 5]. The current study investigates the importance of social media analysis in catastrophic situations such as COVID-19 pandemic. This is extremely difficult to understand a giant amount of text automatically that could lead to better understanding the critical situation and helpful for the policy makers to take appropriate measures and design policies in such situations.

4.1 Analysis of Social Media Texts during Covid-19

Social Media platforms are different in nature and have different characteristics. It means that each platform has a different data organization, data handling and postings flow. This is why it is extremely important to understand how the data is published in these platforms. Most of the platforms, such as Facebook, Twitter accept a variety of data types like pictures, text, videos. The pictures, video types data is beyond the scope of this study. The automatic algorithms divide sentences into words to understand the

text published on social media platforms. This text can be twitter postings, facebook's stories, comments etc.

In addition to this, such types of data pose new issues such as incorrect word spellings, confusing abbreviations that may lead to a completely different meaning. It is extremely important to mention that the performance of supervised learning algorithms is highly data quality dependent. The better the data quality is, the good results can be obtained through supervised learning approaches. Therefore, the correctness of a giant amount of data is a troublesome task and this is why traditional supervised methods are inefficient to provide better results in real time prediction tasks. Similarly, unsupervised approaches [39, 40] use clustering techniques to find word similarities of correct word and its different misspellings. However, in the case of predicting COVID-19, due to its dependency on rigid metrics for similarity that influence the clusters, these methods are also not a good option to deal with catastrophic situations.

Deep learning methods, particularly, predicting the COVID-19 outbreak [42, 43] provides better results by deducting high-level abstractions (e.g. meaning) from lower levels (e.g. letters or substring), and these methods are good in representing words similarities and its various spellings variations. In addition to this, these methods are really good in the case of COVID-19 when the training set is not substantial [48, 49].

5 Discussion

It is evident that recent machine learning and deep learning methods have proven great potential in combating the COVID-19 pandemic. These techniques have been extremely helpful for governments and organizations to process big data related to COVID-19 and design appropriate policies like lockdown, social distancing, finding challenges people face on a daily basis to improve the medical health services. However, a number of challenges are still unaddressed and require the attention of responsible institutions. Although a substantial number of researches have been conducted in creating datasets related to Covid-19 [33, 61, 62, 63, 64, 65], a critical challenge arises from the lack of standard datasets.

As discussed in previous sections, various algorithms have been used to process the data, in particular, for the virus detection. For example, recent studies [63, 64] suggested different algorithms for virus detection. These algorithms are tested on different datasets, the study [63] achieved 82.9% accuracy while the study [64] demonstrated that their algorithms outperformed by getting 98.27% accuracy in detecting virus. Nevertheless, it is crucial to come to a conclusion to decide which algorithm performed other algorithms to detect COVID-19 virus since both datasets are different in size.

Standard datasets can be extremely fruitful to combat COVID-19 prompt and effectively. Substantial researchers used online data, concatenated with some other online available datasets to check their proposed algorithm performance. However, the standard dataset collection issue can be easily addressed if multiple organizations, government, tech firms, and various main health organizations (e.g., WHO and CDC) collectively assemble big datasets. This collaboration will foster not only identifying the solutions prompt, but also help in providing unique sources for the high-quality

research work. To keep this in mind, the study [65] proposed a CORD-19 dataset that is created by multiple organizations.

These organizations are Georgetown's Center for Security, Microsoft Research, National Institutes of Health, and many other companies like Allen Institute for AI, and Chan Zuckerberg Initiative.

6 Questions for Further Research

It is prominent by a number of studies that AI and big data technologies play a significant role in combating the COVID19 pandemic. However, after reviewing the state-of-the-art literature, several research questions arose that motivated us to work on future research. For example, in the future research, we will investigate not only how AI analytics and big data can help us in studying the COVID-19 virus protein structure, but also understand how the performance in terms of accuracy and reliability of the data analytics tools can be improved to develop COVID-19 vaccine. This research will be fruitful in terms of economic and scientific perspectives.

7 Conclusion

The study aims to provide an overview of the state-of-the-art methods used to process and analyze the digital media text in the battle against the COVID-19 pandemic. In the beginning, the study provided an introduction about the outbreak of the COVID-19 disease. Subsequently, the applications of AI in mining digital media text and effective methods used to combat the COVID-19 disease. The study also reviewed and highlighted the limitations of the traditional machine learning models that are unable to provide good performance to track and predict people's response and attitude towards COVID-19 epidemic. On the other hand, a comprehensive analysis of multiple deep learning methods is provided that can be helpful to mitigate the issue and address the catastrophic situation effectively. In addition to this, the study also shedded light on the challenges that require special attention to successfully fight against this havoc. Finally, the study highlighted questions for future research and concluded that the deep learning models such as BERT outperforms other methods in mining digital text related to COVID-19 pandemic.

References

1. Quattrociocchi, W.: Part 2-social and political challenges: 2.1 western democracy in crisis? In: Global Risk Report World Economic Forum (2017)
2. Kim, L., Fast, S.M., Markuzon, N.: Incorporating media data into a model of infectious disease transmission. PloS One, 14(2) (2019)
3. Sharot, T., Sunstein, C.R.: How people decide what they want to know. Nature Human Behaviour, pp. 1–6 (2020)
4. Shaman, J., Karspeck, A., Yang, W., Tamerius, J., Lipsitch, M.: Real-time influenza forecasts during the 2012–2013 season. Nature Communications, 4(1), pp. 1–10 (2013)
5. Viboud, C., Vespignani, A.: The future of influenza forecasts. Proceedings of the National Academy of Sciences, 116(8), pp. 2802–2804 (2019)

6. Kulshrestha, J., Eslami, M., Messias, J., Saptarshi, B.Z., Ghosh., Krishna, P. Gummadi., & Karahalios, K.: Quantifying search bias: Investigating sources of bias for political searches in social media. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing, pages 417–432 (2017)
7. Starnini, M., Frasca, M., Baronchelli, A.: Emergence of metapopulations and echo chambers in mobile agents. Scientific Reports, 6, pp. 31834 (2016)
8. Schmidt, A., Zollo, F., Scala, A., Betsch, C., Quattrociocchi, W.: Polarization of the vaccination debate on facebook. Vaccine, 36(25), pp. 3606–3612 (2018)
9. Schmidt, A.L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., Stanley, H.G., Quattrociocchi, W.: Anatomy of news consumption on facebook. Proceedings of the National Academy of Sciences, 114(12), pp. 3035–3039 (2017)
10. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), pp. 554–559 (2016)
11. Bessi, A., Coletto, M., Davidescu, G.A., Scala, A., Caldarelli, G., Quattrociocchi, W.: Science vs conspiracy: Collective narratives in the age of misinformation. PloS One, 10(2) (2015)
12. Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., Quattrociocchi, W.: Debunking in a world of tribes. PloS One, 12(7) (2017)
13. Baronchelli, A.: The emergence of consensus: a primer. Royal Society Open Science, 5(2), pp. 172189 (2018)
14. Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., Quattrociocchi, W.: Echo chambers: Emotional contagion and group polarization on facebook. Scientific reports, 6, pp. 37825 (2016)
15. Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A.: Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences, 115(37), pp. 9216–9221 (2019)
16. Del Vicario, M., Quattrociocchi, W., Scala, A., Zollo, F.: Polarization and fake news: Early warning of potential misinformation targets. ACM Transactions on the Web (TWEB), 13(2), pp. 1–22 (2019)
17. Wardle, C., Derakhshan, H.: Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe report, 27 (2017)
18. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science, 359(6380), pp. 1146–1151 (2018)
19. Ruths, D.: The misinformation machine. Science, 363(6425), pp. 348–348 (2019)
20. Bovet, A., Makse, H.A.: Influence of fake news in twitter during the us presidential election. Nature Communications, 10(1), pp. 1–14 (2019)
21. Nauhwar, A.: Digital Data: How digital data is going to become a gold for all industries in upcoming year. Iconic Research and Engineering Journals (2020)
22. NodeGraph: How much data is on the internet? The Big Data Facts Update 2020. https://www.nodegraph.se/how-much-data-is-on-the-internet/ (2020)
23. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159–190 (2006)
24. Amjad, M., Sidorov, G., Zhila, A.: Data augmentation using machine translation for fake news detection in the Urdu language. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, pp. 2537–2542 (2020)
25. Amjad, M., Sidorov, G., Zhila, A., Gomez-Adorno, H., Voronkov, I., Gelbukh, A.: Bend the truth: 2541 a benchmark dataset for fake news detection in Urdu language and its evaluation. Journal of Intelligent & Fuzzy Systems, In Press (2020)
26. Rish, I.: An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop On Empirical Methods In Artificial Intelligence, 3(22), pp. 41–46 (2001)

27. Amjad, M., Voronkov, I., Saenko, A., Gelbukh, A.: Comparison of text classification methods using deep learning neural networks. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing) (2019)
28. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R.K.: Speech emotion recognition using support vector machine (2020)
29. Shi, F., Xia, L., Shan, F., Wu, D., Wei, Y., Yuan, H., Shen, D.: Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification (2020)
30. Chen, C., Yan, J.T., Zhou, N., Zhao, J.P., Wang, D.W.: Analysis of myocardial injury in patients with COVID-19 and association between concomitant cardiovascular diseases and severity of COVID-19. Zhonghua Xin Xue Guan Bing Za Zhi, 48 (2020)
31. Zhong, B.L., Luo, W., Li, H.M., Zhang, Q.Q., Liu, X.G., Li, W.T., Li, Y.: Knowledge, attitudes, and practices towards COVID-19 among Chinese residents during the rapid rise period of the COVID-19 outbreak: a quick online cross-sectional survey. International Journal of Biological Sciences, 16(10), pp. 1745 (2020)
32. Forrester, J.D., Nassar, A.K., Maggio, P.M., Hawn, M.T.: Precautions for operating room team members during the COVID-19 pandemic. Journal of the American College of Surgeons (2020)
33. Chen, E., Lerman, K., Ferrara, E.: Covid-19: The first public coronavirus twitter dataset (2020)
34. Park, H.W., Park, S., Chong, M.: Conversations and medical news frames on twitter: infodemiological study on COVID-19 in South Korea. Journal of Medical Internet Research, 22(5), e18897 (2020)
35. Clerkin, K.J., Fried, J.A., Raikhelkar, J., Sayer, G., Griffin, J.M., Masoumi, A., Schwartz, A.: Coronavirus disease 2019 (COVID-19) and cardiovascular disease. Circulation (2020)
36. Roda, W.C., Varughese, M.B., Han, D., Li, M.Y.: Why is it difficult to accurately predict the COVID-19 epidemic?. Infectious Disease Modelling (2020)
37. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning. IEEE Transactions on Neural Networks, 20(3), pp. 542–542 (2009)
38. Pung, R., Chiew, C.J., Young, B.E., Chin, S., Chen, M.I., Clapham, H.E., Low, M.: Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. The Lancet (2020)
39. Hisada, S., Murayama, T., Tsubouchi, K., Fujita, S., Yada, S., Wakamiya, S., Aramaki, E.: Syndromic surveillance using search query logs and user location information from smartphones against COVID-19 clusters in Japan (2020)
40. Hu, Z., Ge, Q., Jin, L., Xiong, M.: Artificial intelligence forecasting of COVID-19 in China (2020)
41. Jahanbin, K., Rahmanian, V.: Using twitter and web news mining to predict COVID-19 outbreak. Asian Pacific Journal of Tropical Medicine, 13 (2020)
42. Ferrara, E.: COVID-19 on twitter: bots, conspiracies, and social media activism (2020)
43. Alqudah, A.M., Qazan, S., Alqudah, A.: Automated systems for detection of COVID-19 using chest x-ray images and lightweight convolutional neural networks (2020)
44. Jelodar, H., Wang, Y., Orji, R., Huang, H.: Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach (2020)
45. Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Cao, K.: Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology (2020)
46. Alam, F., Shaar, S., Nikolov, A., Mubarak, H., Martino, G.D.S., Abdelali, A., Nakov, P.: Fighting the COVID-19 Infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society (2020)

47. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter (2020)
48. Posadas-Durán, J.P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D.: Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing 21(3), pp. 627–639 (2017)
49. Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., Pinto, D.: Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing 100 (7), pp. 741–756 (2018)
50. Bengio, Y., LeCun, Y.: Scaling learning algorithms towards AI. Large-Scale Kernel Machines, 34(5), pp. 1–41 (2007)
51. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: A review. Neurocomputing, 187, pp. 27–48 (2016)
52. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Geessink, O.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), pp. 2199–2210 (2017)
53. Bengio, Y., Goodfellow, I., Courville, A.: Deep learning. 1, MIT press (2017)
54. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML Workshop On Unsupervised and Transfer Learning, pp. 17–36 (2017)
55. Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Iyengar, S.S.: A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5), pp. 1–36 (2018)
56. Sun, W., Zheng, B., Qian, W.: Computer aided lung cancer diagnosis with deep learning algorithms. In: Medical imaging 2016: computer-aided diagnosis. 9785, pp. 97850Z, International Society for Optics and Photonics (2016)
57. Zhang, N., Zhang, R., Yao, H., Xu, H., Duan, M., Xie, T., Zhou, F.: Severity Detection For the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. (2020)
58. Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Shi, J., He, G.: Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua, 63(1), pp. 537–551 (2020)
59. Albahri, A.S., Hamid, R.A.: Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. Journal of Medical Systems, 44(7) (2020)
60. Pham, Q.V., Nguyen, D.C., Hwang, W.J., Pathirana, P.N.: Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts (2020)
61. Fong, S.J., Li, G., Dey, N., Crespo, R.G., Herrera-Viedma, E.: Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak (2020)
62. Gao, Z., Yada, S., Wakamiya, S., Aramaki, E.: Naist covid: multilingual COVID-19 twitter and weibo dataset (2020)
63. Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Xu, B.: A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19), MedRxiv (2020)
64. Ozkaya, U., Ozturk, S., Barstugan, M.: Coronavirus (COVID-19) classification using deep features fusion and ranking technique (2020)
65. Kaggle: COVID-19 open research dataset challenge (CORD-19): an AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House. www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (2020)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx