Alvaro David Orjuela-Cañón
Jesus A Lopez
Oscar J Suarez (Eds.)

Communications in Computer and Information Science 2212

Applications of Computational Intelligence

7th IEEE Colombian Conference, ColCACI 2024
Pamplona, Colombia, July 17–19, 2024
Revised Selected Papers

Springer

Communications in Computer and Information Science 2212

Rationale

The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics

The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors

Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (http://link.springer.com) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process

The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing

CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start

To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.

Alvaro David Orjuela-Cañón · Jesus A. Lopez ·
Oscar J. Suarez
Editors

Applications of Computational Intelligence

7th IEEE Colombian Conference, ColCACI 2024
Pamplona, Colombia, July 17–19, 2024
Revised Selected Papers

Springer

Editors
Alvaro David Orjuela-Cañón
Universidad del Rosario
Bogotá D.C., Colombia

Jesus A. Lopez
Universidad Autónoma de Occidente
Cali, Colombia

Oscar J. Suarez
Universidad de Pamplona
Pamplona, Colombia

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-88853-3 ISBN 978-3-031-88854-0 (eBook)
https://doi.org/10.1007/978-3-031-88854-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Computational intelligence (CI), an important area related to artificial intelligence, provides solutions that are increasingly employed in engineering problems in different areas in the Latin America (LA) region. In the present version of the Applications of Computational Intelligence book different problems are addressed, including solar and photovoltaic forecasting, microseismical signal analysis, LLM performance analysis, evolution in translation systems, recognition of actors and peacebuilding, control in autonomous vehicles, and electroencephalography signals analysis, mainly. These problems are commonly addressed in this part of the world, and these applications have great potential in developing countries.

This seventh edition contains the best extended and selected revised papers from the 2024 Colombian Conference on Applications of Computational Intelligence (ColCACI 2024), which was fortified with contributions from scientists, engineers and practitioners working on both applications and theory of CI techniques. We received 35 papers by authors from countries such as: Colombia, Ecuador, Spain, Mexico, USA, Honduras, Argentina, Canada, and Peru. The conference had 22 oral presentations applying in-person mode and also other ways to share ideas and talk about CI topics. Thus, the event was an international forum for CI researchers and practitioners to share their most recent advancements and results. This book contains the post-proceedings with revised selected papers, including the best 11 papers presented, as extended versions of work exhibited at the conference. We are committed to continue working to offer excellent IEEE ColCACI events in future.

Finally, we would like to thank the IEEE Colombia section, the IEEE Computational Intelligence Society Colombian chapter, Universidad Autónoma de Occidente, Universidad del Rosario, and Universidad de Pamplona for hosting the conference. In addition, we would like to thank Springer for their support and all reviewers that did such awesome work. Also, special thanks to all the volunteers, participants, and the whole crew who worked together to create a successful conference.

December 2024

Alvaro David Orjuela-Cañón
Oscar J. Suárez
Jesus A. Lopez

Organization

Organizers

General Chair

Oscar Javier Suarez Sierra	Universidad de Pamplona, Colombia

Technical Chair

Jesús Alfonso López Sotelo	Universidad Autónoma de Occidente, Colombia

Publication Co-chairs

Alvaro David Orjuela-Cañón	Universidad del Rosario, Colombia
Carlos Jesus Vega Perez	Uptime Analytics, Colombia
José Luis Paniagua Jaramillo	Universidad Autónoma de Occidente, Colombia

Financial Chair

Alvaro David Orjuela-Cañón	Universidad del Rosario, Colombia

Program Committee

Alvaro David Orjuela Cañón	Universidad del Rosario, Colombia
Jesus A. Lopez	Universidad Autónoma de Occidente, Colombia
Julián David Arias Londoño	Universidad de Antioquia, Colombia
Juan Carlos Figueroa García	Universidad Distrital Francisco José de Caldas, Colombia
Danton Ferreira	Universidade Federal de Lavras, Brazil
Efren Gorrostieta	Universidad Autónoma de Queretaro, Mexico
Cristian Rodríguez Rivero	UC Davis, USA
Jose Alfredo Costa	Universidade Federal do Rio Grande do Norte, Brazil
Javier Mauricio Antelis	Instituto Tecnológico de Monterrey, Mexico
Leonardo Forero Mendoza	Universidade Estadual do Rio de Janeiro, Brazil
Carmelo Bastos Filho	Universidade de Pernambuco, Brazil

Edgar Sánchez	CINVESTAV, Unidad Guadalajara, Mexico
Guilherme Alencar Barreto	Universidade Federal do Ceará, Brazil
Gonzalo Acuña Leiva	Universidad de Santiago de Chile, Chile
Carlos Alberto Cobos Lozada	Universidad del Cauca, Colombia
Juan Bernardo Gómez Mendoza	Universidad Nacional de Colombia Sede Manizales, Colombia
Diego Peluffo Ordóñez	Universidad Técnica del Norte, Ecuador
Gerardo Muñoz Quiñones	Universidad Distrital Francisco José de Caldas, Colombia
Jorge Eliécer Camargo Mendoza	Universidad Antonio Nariño, Colombia
Claudia Victoria Isaza Narvaez	Universidad de Antioquia, Colombia
Sandra Esperanza Nope Rodríguez	Universidad del Valle, Colombia
Jesús Alfonso López Sotelo	Universidad Autónoma de Occidente, Colombia
Cesar Hernando Valencia Niño	Universidad Santo Tomás, Colombia
Miguel Melgarejo Rey	Universidad Distrital Francisco José de Caldas, Colombia
Wilfredo Alfonso Morales	Universidad del Valle, Colombia
Diana Consuelo Rodríguez	Universidad del Rosario, Colombia
Humberto Loaiza	Universidad del Valle, Colombia
Eduardo Francisco Caicedo Bravo	Universidad del Valle, Colombia
Alexander Molina Cabrera	Universidad Tecnológica de Pereira, Colombia
Luiz Pereira Caloba	Universidade Federal de Rio de Janeiro, Brazil
Alvaro Gustavo Talavera	Universidad del Pacífico, Peru
Efraín Mayhua-López	Universidad Católica San Pablo, Peru
Yván Tupac	Universidad Católica San Pablo, Peru
Ana Teresa Tapia	Escuela Superior Politécnica del Litoral, Ecuador
Miguel Núñez del Prado	Universidad del Pacífico, Peru
Heitor Silvério Lopes	Universidade Tecnológica Federal de Paraná, Brazil
Waldimar Amaya	ICFO - Institute of Photonic Sciences, Spain
Leonardo Franco	Universidad de Málaga, Spain
Carlos Andrés Peña	University of Applied Sciences Western Switzerland, Switzerland
Edwin Alexander Cerquera	University of Florida, USA
Nadia Nedjah	Universidade Estadual do Río de Janeiro, Brazil
María Daniela López de Luise	CI2S Lab, Argentina
Gustavo Eduardo Juárez	Universidad Nacional de Tucumán, Argentina

Contents

A Deep Learning Framework for the Automatic Classification of Microseisms at Cotopaxi Volcano

Román Lara-Cueva[1](✉), Iván Iglesias[1], Alejandro Rosero[1], Diego Benítez[2], Noel Pérez-Pérez[2], and José Luis Rojo-Álvarez[3]

[1] Grupo de Investigación en Sistemas Inteligentes (WiCOM-Energy) and Ad Hoc Networks Research Center (CIRAD), Departamento de Eléctrica, Electrónica y Telecomunicaciones, Universidad de Las Fuerzas Armadas ESPE, Sangolquí 171103, Ecuador
{ralara,larosero4}@espe.edu.ec, ivaniglesias@ieee.org

[2] Colegio de Ciencias e Ingenierías "El Politécnico", Universidad San Francisco de Quito USFQ, Quito 170157, Ecuador
{dbenitez,nperez}@usfq.edu.ec

[3] Department of Signal Theory and Communications Telematics, and Computing Systems, Universidad Rey Juan Carlos, 28933 Madrid, Spain
joseluis.rojo@urjc.es

Abstract. This paper presents a comprehensive framework for an automated classification system of Long Period (LP) and Volcano Tectonic (VT) seismic events from the Cotopaxi volcano, employing two prominent Deep Learning models: Stacked Autoencoders and Deep Neural Networks. The dataset was provided by the Instituto Geofísico de la Escuela Politécnica Nacional (IGEPN), which consists of 1,044 LP events and 101 VT events. To enhance the model's performance, data augmentation techniques were applied, generating synthetic microseismic data using a Conditional Generative Adversarial Network (CGAN), ultimately expanding the dataset to 10,000 microseisms, comprising both real and synthetic samples. In the feature extraction stage, we calculated the Power Spectral Density (PSD) using Welch's method and a 512-point Fast Fourier Transform (FFT), resulting in 257 frequency-based features. Additionally, wavelet coefficients were obtained via Discrete Wavelet Transform (DWT) using the Symlet wavelet with six levels of decomposition. PSD was also computed on these wavelet coefficients, producing 257 additional frequency-scale features. The enriched feature set was used to train the Deep Learning models, with performance evaluated using standard classification metrics. The Stacked Autoencoder model outperformed others, achieving an accuracy of 99.73% and a balanced error rate (BER) of 0.0027, corresponding to just 3 errors per 1,000 classified microseisms. These results substantially exceed the geophysical accuracy requirement of a BER below 0.01. This paper demonstrates the efficacy of Deep Learning models in distinguishing between LP and VT events, offering new opportunities for enhancing volcanic monitoring and hazard assessment strategies.

Keywords: Classification System · Cotopaxi Volcano · Deep Learning · Supervised Learning · Volcanic Microseisms

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 1–13, 2025.
https://doi.org/10.1007/978-3-031-88854-0_1

1 Introduction

Seismology plays a critical role in mitigating volcanic hazards by enabling the detection and classification of microseisms, which are vital for forecasting potential eruptions and issuing early warnings to at-risk populations. This proactive approach helps protect lives in vulnerable communities [1, 2]. Traditionally, volcanic monitoring organizations perform this task semi-automatically, where detection often relies on the STA/LTA (Short-Term Average/Long-Term Average) algorithm, which achieves an accuracy of approximately 93%. Classification, however, is typically conducted manually by experts, making it subjective and labor-intensive, especially during periods of increased volcanic activity, where large amounts of seismic data can accumulate [3].

In recent years, Machine Learning (ML) has emerged as a powerful tool to enhance both detection and classification processes. ML has been particularly effective in boosting the performance of the STA/LTA algorithm, improving detection accuracy by at least 4% [4–8]. For classification, techniques such as k-nearest neighbors (k-NN), support vector machines (SVM), decision trees (DT), and neural networks (NN) have been successfully applied, achieving classification accuracies greater than 93% [9–12]. More recently, Deep Learning (DL) methods have been explored, using a two-dimensional (2-D) approach by treating microseism spectrograms as images. This has resulted in detection accuracies of between 78% and 99% in distinguishing microseisms from background noise (BN), representing significant progress [13, 14]. Furthermore, DL has also achieved a classification accuracy of up to 97% in both binary and multiclass tasks [15–17]. However, an important metric for geophysical and volcanological observatories is the balanced error rate (BER), which must be below 0.01—meaning no more than 1 misclassified microseism per 100. To date, this stringent requirement has yet to be consistently met in previous studies.

This paper aims to develop DL algorithms to automate the classification of Long Period (LP) and Volcano Tectonic (VT) microseisms from the Cotopaxi volcano. Specifically, we propose to evaluate two DL models: Deep Neural Networks (DNN) and Stacked Autoencoders (SA) [18]. The framework includes multiple stages, such as microseism processing, feature extraction using Power Spectral Density (PSD), and coefficients from the Discrete Wavelet Transform (DWT). These extracted features are then processed by the DL classifiers. Finally, the system's performance is assessed using key evaluation metrics pertinent to seismic classification.

2 Proposed Methodology

Our proposal presents a one-dimensional (1-D) binary classification system for volcanic microseisms using DL algorithms. The block diagram presented (see Fig. 1) illustrates the stages that compose the operation of the proposed system. In the first stage, three databases are described, the first one consisting of 1,045 volcanic microseisms between LP and VT signals that we denominated 'Real DB', another that has 10,000 microseisms generated artificially by CGAN algorithm that we named 'Synthetic DB' and finally 10,000 volcanic microseisms, a mix of real and synthetic microseisms of the LP and VT types, and that is called 'Mixed DB'. Then, in the signal processing stage, a finite impulse response (FIR) band-pass filter is applied to each microseism, with cutoff frequencies between 1 and 50 Hz, and its amplitude is normalized between 0 and 1. In the feature extraction stage, the spectral estimation of the PSD is obtained, as well as the approximation and detail coefficients using the discrete wavelet transform, which allows for a database composed of 514 features. Subsequently, in the classification stage, two DL models are employed: the SA and the DNN, which consist of multiple layers of elementary blocks. Finally, the proposed models are evaluated on the basis of performance metrics.

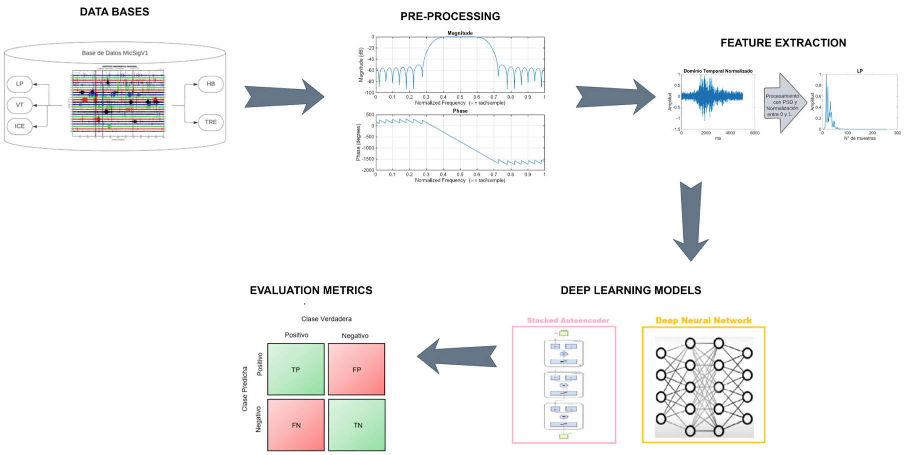


Fig. 1. Proposed DL pipeline.

2.1 Database

The databases used consists of a mix of real and synthetic microseisms, comprising volcanic microseisms LP and VT. The real microseisms consist of 1,044 LP and 101 VT microseisms extracted from the MicSigV1 database, which is available from the Instituto Geofísico de la Escuela Politécnica Nacional (IGEPN) [19]. Synthetic microseisms are generated using a Conditional Generative Adversarial Network (CGAN) model presented in [20], producing 5,000 LP and 5,000 VT, which are used to improve classification results based on the data augmentation process [21, 22]. Finally, the database consists of a random mixture of unprocessed real and synthetic microseisms, totaling M = 10,000 microseisms (5,000 LP and 5,000 VT), including all real microseisms complemented with synthetic ones, represented by the matrix: $\boldsymbol{S} = \{s_1^T, s_2^T, \cdots, s_M^T\}^T$ where s_i contains a microseism.

To address the class imbalance, present in the original dataset provided by IGEPN, which contains 1044 LP signals and only 101 VT signals, we employed the k-folds cross-validation method. This approach was implemented to ensure a balanced distribution of data for training and evaluating DL models, specifically the Stacked SA and DNN.

The k-folds method involves partitioning the data into k subsets, or folds, and iteratively using each fold as a test set while the remaining k-1 folds serve as the training set.

In our specific application, the method was adapted to select 101 LP signals from different intervals within the dataset to match the quantity of VT signals. This balancing technique is crucial for mitigating bias and enhancing the generalization capability of the DL models. By ensuring an equal representation of both signal types during the training and testing phases, we aim to improve the models' performance and robustness in distinguishing between LP and VT signals.

2.2 Processing

Spectral analysis of real microseisms reveals a prominent spectral component around 0.2 Hz, corresponding to the impact of the ocean on the continent [23], which is filtered out because it overshadows other components of interest for our analysis [24]. Each microseismical event has a frequency range, but it varies between events. The theoretical basis indicates that it is better to analyze the signals from 1 Hz to avoid eliminating important information from the microseisms [2]; therefore, the signals were filtered using a FIR band-pass filter, non-causal, zero-phase, order 256, and cutoff frequencies defined at 1 Hz and 50 Hz [25]. Finally, the dataset is normalized to an amplitude represented on the y-axis between 0 and 1, applying the operator $g_h(.)$, which represents the application of the filtering and normalization stages to each s_i, producing feature vectors $h_i = g_h(s_i)$, and then generating the matrix: $\mathbf{H} = \{h_1^T, h_2^T, \cdots, h_M^T\}^T$.

2.3 Feature Extraction

In previous traditional ML works, the focus has been on using features in the time, frequency, and scale domains, while for DL, the focus has been on using 2-D spectrograms with pre-trained networks such as Alexnet, Squeezenet, and Googlenet [14, 26], which work with low-resolution images. Our approach proposes evaluating DL techniques with 1-D frequency and scale domain features using PSD and the PSD of the wavelet transform coefficients. This method significantly reduces training times and time processing.

Power Spectral Density. The PSD represents the power contained in a signal per unit of frequency. It is a fundamental tool in signal processing and analysis, widely used by geophysical institutes, since the PSD helps to understand the dynamics of seismic waves and provides valuable information for hazard assessment [27]. The power spectral estimator used is Welch, due to it being a technique frequently used by volcanological institutes and has been evaluated in [28], where the Welch and Burg estimators present the best results. A medium resolution of 512 points of the Fast Fourier Transform (FFT) was defined for the spectral estimation of the volcanic microseisms' PSD, obtaining 257 frequency features. The operator $g_f(.)$ represents the application of the PSD feature extraction stage to each h_i, which produces feature vectors $f_i = g_f(h_i)$, and then the matrix $\boldsymbol{F} = \{f_1^T, f_2^T, \cdots, f_M^T\}^T$ is generated.

Wavelet Transform Coefficients. The Fourier transform is a powerful tool for analyzing the components of a stationary signal. However, it fails when analyzing non-stationary signals, while the wavelet transform allows analyzing the components of a non-stationary signal. This capability makes the Wavelet Transform especially valuable in volcanic signal applications, as it serves to decompose signals into their frequency components, revealing details not visible with the Fourier Transform [29]. The continuous Wavelet Transform (CWT) faces a data redundancy problem when handling large amounts of data, while the DWT overcomes this problem by using orthogonal wavelets. Therefore, in this paper, two types of the discrete wavelet were used for coefficient extraction. In order to obtain the optimal parameters, a mother wavelet of type 10, known as 'DB10' in the case of the Daubechies model and 'SYM10' for the Symlet model, must be used. The detail coefficients of levels 2, 3, 4 and 5 were obtained, in addition to the level 6 approximation coefficient, after which the coefficients obtained were unified in a single vector, both for Daubechies and Symlets [30], allowing observation of frequencies of interest, as showed below in Fig. 2. Finally, the PSD of the wavelet coefficients was obtained, providing detailed information on the spectral distribution of a signal in terms of scale and time location, useful for analyzing microseisms and using these features on different time scales, applying the operator $g_w(.)$, which represents the application of the Wavelet Transform coefficient extraction stage to each h_i, producing feature vectors $w_i = g_w(h_i)$, and then enerating the matrix $\boldsymbol{W} = \{w_1^T, w_2^T, \cdots, w_M^T\}^T$. Subsequently, Welch's PSD estimator is applied to the Wavelet Transform coefficients, with 512 points of the FFT, obtaining 257 scale-frequency features, applying the operator $g_f(.)$, which represents the application of the PSD to each e_i, producing feature vectors $e_i = g_f(w_i)$, and then generating the matrix $\boldsymbol{E} = \{e_1^T, e_2^T, \cdots, e_M^T\}^T$. Finally, the frequency features (257) and the features obtained by applying the PSD to the wavelet

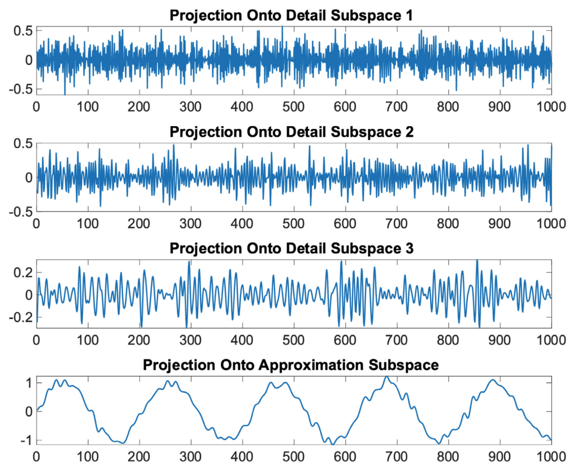


Fig. 2. Decomposition and coefficients of the wavelet transform.

(257) are concatenated, forming a database with 514 features, generating the matrix $\boldsymbol{B} = [\boldsymbol{F} : \boldsymbol{E}]$.

In this article, two DL techniques applicable to our 1-D matrix **B** are evaluated, selecting the neural networks: SA and DNN.

2.4 Deep Learning Models

In this article, two DL techniques applicable to our 1-D matrix **B** are evaluated, selecting the neural networks: SA and DNN.

Stacked Autoencoder. An autoencoder is a neural network that reduces dimensionality and learns data features. It consists of an encoder and a decoder function that produces a reconstruction, with a hidden layer serving as the code to represent the input. However, if the network learns to copy the input perfectly, it is not particularly useful. Therefore, autoencoders are often designed to prioritize which aspects of the input to copy, often learning useful properties of the data in the process. To avoid such perfect copying, autoencoders often have constraints that allow them to copy only approximately, and only the input that resembles the training data, which can be especially useful in generation and classification applications [31]. By trial and error, the values that maximize BER were identified, including the number of neurons, the number of epochs, the weight of L2 regularization, the regularization of sparsity, and the proportion of sparsity. For more information, it is suggested to review [31–33].

DNN. DNNs consist of multiple layers of building blocks, such as nonlinear activation layers or more complex subnets. Each layer uses the output of the previous one to learn a gradual hierarchical representation. The learning is performed using the backpropagation algorithm (BP), which adjusts the parameters to minimize the output error. DNNs have advanced in areas such as image, video, speech, and audio processing, enabling the extraction of complex and abstract features [34]. The term "deep" derives from the fact that neural networks have multiple hidden layers. These layers are important because they allow the network to learn increasingly complex representations of the input. The more layers the network has, the more abstract features it can learn. However, the training process can be computationally expensive [35]. By trial and error, the values that maximize the BER were identified, including the number of hidden layers, the number of output neurons, and the mini-batch size. For more information, it is suggested to review [35, 36].

To determine the optimal signal distribution from datasets for DL model development, a series of experiments was conducted with varying allocation ratios. The examined splits included 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, and 20:80 for model training and evaluation, respectively. Moreover, these proportions have been validated in previous research by the team, encompassing both ML and DL model development.

2.5 Performance Metrics

The performance evaluation of the classifiers is conducted in terms of accuracy (A), precision (P), recall (R), specificity (S), and BER [37].

$$\mathrm{A}(\%) = \frac{\mathrm{TP} - \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100 \tag{1}$$

$$\mathrm{P}(\%) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}} \times 100 \tag{2}$$

$$\mathrm{R}(\%) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \tag{3}$$

$$\mathrm{S}(\%) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100 \tag{4}$$

$$\mathrm{BER} = 1 - \left(\frac{\mathrm{R} + \mathrm{S}}{200}\right) \tag{5}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. We calculate these performance measures for each DL model.

3 Results

In this section, we present the results obtained by applying the proposed methodology to our Cotopaxi volcano dataset, ensuring the independence of all data points. The experiments were conducted on a computer with an 8th generation Intel Core i7 processor, 16 GB of RAM, and Matlab® R2022a.

Feature Extraction. DNNs The result of applying Welch's PSD estimator to each volcanic microseism is shown (see Fig. 3), which was applied to all 10,000 microseisms: original and synthetic. As can be seen, each microseism is transformed into the frequency domain, obtaining 257 features. All records were normalized, as defined in the previous section. When applying the DWT with the Symlet mother wavelet to the volcanic microseisms and extracting the approximation and detail coefficients explained in the previous section and then applying the PSD to these vectors, an additional 257 features were obtained, as presented (see Fig. 4).

As shown (see Fig. 5), the PSD obtained from the microseisms, as well as the PSD of the wavelet coefficients, were concatenated to finally obtain the matrix **B** of 514 features.

Deep Learning Models. Our proposed methodology was applied to matrix **B** to select the optimal parameters of the proposed models based on their performance metrics. In each model, the parameters that maximize performance metrics were identified through a trial-and-error experimentation process. The following sections of the document introduce the tests conducted with the algorithms and their corresponding results.

In order to determine the appropriate training/testing data proportions, tests were conducted varying these proportions, the best results were obtained with a 70/30 partition.

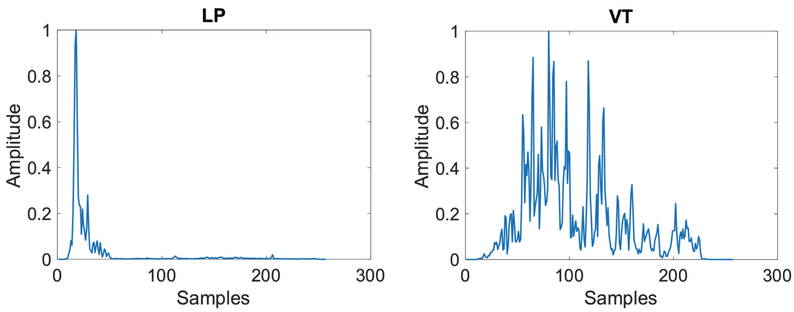


Fig. 3. Representation of the PSD of Cotopaxi LP and VT microseisms.

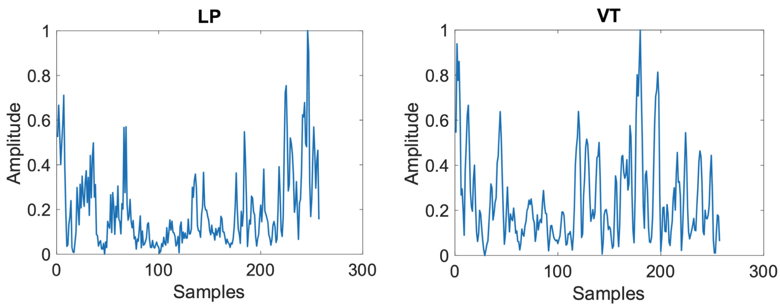


Fig. 4. Representation of the PSD of the coefficients of the Wavelet transform of the LP and VT microseisms of Cotopaxi.

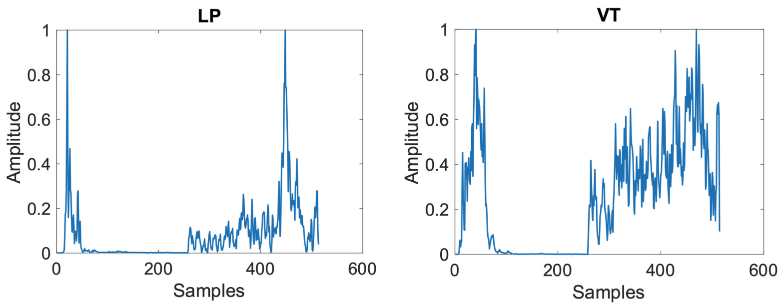


Fig. 5. Concatenated PSD representation with the PSD of the wavelet transform coefficients of LP and VT microseisms from the Cotopaxi volcano.

The optimal parameters for each model that maximized performance metrics with this partition were identified. For SA, the following parameters were identified: 104 neurons, 300 epochs, L2 regularization weight of 0.0060, a sparsity regularization of 7.1340, and a sparsity proportion of 0.3000. Meanwhile, for the DNN model, the parameters were: hidden layers of 50 neurons, output layer with 2 neurons, and a mini-batch size of 16.

These parameters were found under the consideration that the BER value should be below 0.01.

Four tests of the k-folds method are applied on the training data for the unbalanced 'Real' database, also without applying the wavelet transform into that data, and in all tests very similar performance metrics were obtained for both DL algorithms, with the above mentioned, the results of this test are presented in Table 1.

Table 1. Tests applying k-folds technique in order to mitigate unbalanced data bias.

LP Signals interval	A (%)	P (%)	R (%)	S (%)	BER
[0–101]	96.67	93.33	93.75	99.93	0.0312
[300–401]	98.33	96.67	96.77	99.95	0.0165
[600–701]	99.75	99.53	96.69	96.73	0.0231
[900–1001]	98.33	96.67	96.77	99.95	0.0165

The best general results of the performance metrics after balanced the database and applied both wavelet transforms, Daubechies and Symlets to the microseisms are shown in Table 2.

Table 2. Development metrics for DL Models.

DB Type	Applied Wavelet	Model	A (%)	P (%)	R (%)	S (%)	BER
Real	N/A	SA	98.33	96.70	96.74	99.97	0.0165
Real	N/A	DNN	98.32	96.64	96.80	99.93	0.0164
Synthetic	N/A	SA	99.80	99.93	98.93	99.66	0.0071
Synthetic	N/A	DNN	99.73	99.60	98.60	99.86	0.0077
Mixed	N/A	SA	99.53	99.66	99.66	99.40	0.0047
Mixed	N/A	DNN	99.43	99.33	99.33	99.53	0.0057
Mixed	Daubechies	SA	99.60	99.80	99.79	99.40	0.0040
Mixed	Daubechies	DNN	99.30	99.81	99.80	99.80	0.0070
Mixed	Symlets	SA	99.73	99.53	99.53	99.93	0.0027
Mixed	Symlets	DNN	99.76	99.53	99.86	99.53	0.0035

Obtained Results. As it is shown in Table 2, the best results with the SA model applied into the 'Real' database are an accuracy of 98.33%, a precision of 96.70%, recall of 96.74%, specificity of 99.97% and a BER of 0.0165. With the test for the DNN model applied instead into the 'Real' data we obtained an accuracy of 98.32%, a precision of 96.64%, recall of 96.80%, specificity of 99.93% and a BER of 0.0164. Then, applying the DNN model into 'Synthetic' database the results are an accuracy of 99.73%, a precision of 99.60%, recall of 98.60%, specificity of 99.86% and a BER of 0.0077. With the test for the SA model applied instead into the 'Synthetic' data we obtained an accuracy of 99.80%, a precision of 99.93%, recall of 98.93%, specificity of 99.66% and a BER of 0.0071. With the 'Mixed' database we've got an accuracy of 99.53%, a precision of 99.66%, recall of 99.66%, specificity of 99.40% and a BER of 0.0047 for the SA model applied. With the DNN model applied into the 'Mixed' data we've got an accuracy of 99.43%, a precision of 99.33%, recall of 99.33%, specificity of 99.53% and a BER of 0.0057. For the DNN model applied to the data with a mother Daubechies wavelet we have an accuracy of 99.30%, a precision of 99.81%, recall of 99.80%, specificity of 99.80% and a BER of 0.0070. In the case of the SA model applied to data processed with the wavelet Daubechies transform, an accuracy of 99.60%, an accuracy of 99.80%, a recall of 99.79%, a specificity of 99.40% and a BER of 0.0040 are observed. These parameters indicate excellent results in the classification of volcanic microseisms for both models, with results favoring SA over DNN. However, because the results obtained are slightly inferior to the results with Symlets, the latter will be chosen as the best overall results for microearthquakes of the Cotopaxi volcano. The SA model for the data applied the Symlets mother wavelet presents an accuracy of 99.73%, a precision of 99.53%, recall of 99.53%, specificity of 99.93%, and a BER of 0.0027. For the DNN model, it presents an accuracy of 99.76%, a precision of 99.53%, recall of 99.86%, specificity of 99.53%, and a BER of 0.0035 for data with Symlets' wavelet applied. The results obtained are excellent for the binary classification of volcanic microseisms, with the results slightly favoring DNN over SA. Comparing the results obtained in previous works such as [12, 38–40], where a minimum and maximum accuracy of 94% and 97% and a BER of 0.03 were obtained. Our results show that the binary classifier would misinterpret approximately 3 out of 1,000 events, which exceeds the IGEPN requirement of misinterpreting 1 out of 100 events.

4 Conclusions and Future Work

In the results of the SA and DNN classification models were successfully compared, providing valuable information on their performance in classifying volcanic microseisms. The results show that SA achieved an accuracy of 99.80%, precision and recall of 99.53%, specificity of 99.93%, and a BER of 0.0027, outperforming DNN, which obtained an accuracy of 99.76%, precision and specificity of 99.53%, recall of 99.86%, and a BER of 0.0035. SA showed better results in the classification of LP and VT, although both methods exceeded the IGEPN expectation of a BER below 0.01. After applying the same methodological strategy with the SA and DNN models trained solely on the **F** matrix, containing 1,044 LP and 101 VT volcanic microseisms, which is unbalanced, results below 85.00% in all measured performance metrics and a BER greater than 0.1500 were obtained for all experiments conducted. After the k-folds method were applied in order to balance the database, results improve up to 96.00% mean for all performance metrics measures, and the BER reduces to 0.0165 as max value. Also, the results observed with both wavelet transforms applied, Daubechies and Symlets are excellent for a binary classifier, improving the results treating the data with them. This proves that the use of feature extraction strategies to train and test the classifiers, and the data augmentation procedure to have a balanced and larger dataset for applying DL techniques significantly improved performance metrics by at least 14 percentage points. A microseism classifier system was successfully implemented for the Cotopaxi volcano using DL techniques. By evaluating metrics such as accuracy, precision, specificity, sensitivity, and BER, the effectiveness and reliability of the classifiers were assessed. Both the SA and DNN models exceeded the expectations set by the IGEPN for a BER below 0.01. These results contribute to the field of volcanic seismology, providing a valuable tool to improve the classification of volcanic microseisms at Cotopaxi.

Our research group is interested in continuing work in this line, applying the proposed methodology to publicly available databases such as that of the Llaima Volcano, and working with multiclass classifiers to validate and generalize the proposed methodology.

Acknowledgments. The authors gratefully acknowledge the contribution of Universidad de las Fuerzas Armadas ESPE for the economic support in the development of this project under Research Grants 2013-PIT-014 and 2016-EXT-038, and in part by the Universidad San Francisco de Quito (Poli-Grants Program) under Grants 10100, 12494, 16916 and 24275.

References

1. McNutt, S.R.: Volcano seismology and monitoring for eruptions. International Geophysics Series, vol. 81, no. A, pp. 383–406, 2002
2. Chouet, B.A.: Long-period volcano seismicity: its source and use in eruption forecasting. Nature **380**(6572), 309–316 (1996). https://doi.org/10.1038/380309a0
3. Forte, P., et al.: Volcano monitoring in Latin America: taking a step forward: preface to special issue on volcano observatories in Latin America. Volcanica **4**(S1), 7–33 (2021). https://doi.org/10.30909/vol.04.S1.viixxxiii
4. Bhatti, S.M., et al.: Automatic detection of volcano-seismic events by modeling state and event duration in hidden markov models. J. Volcanol. Geoth. Res. **324**, 134–143 (2016). https://doi.org/10.1016/j.jvolgeores.2016.05.015
5. Lara-Cueva, R., Benítez, D., Carrera, E., Ruiz, M., Rojo-Álvarez, J.: Feature selection of seismic waveforms for long period event detection at cotopaxi volcano. J. Volcanol. Geoth. Res. **316**, 34–49 (2016)
6. Lara-Cueva, R.A., Moreno, A.S., Larco, J.C., Benítez, D.S.: Real-time seismic event detection using voice activity detection techniques. IEEE J. Sel. Top. Appl. Earth Observations Remote Sens. **9**(12), 5533–5542 (2016)
7. Proaño, E., Benítez, D.S., Lara-Cueva, R., Ruiz, M.: On the use of variational mode decomposition for seismic event detection, pp. 1–6 (2018)
8. Carrera, E.V., Pérez, A., Lara-Cueva, R.: Automated systems for detecting volcano-seismic events using different labeling techniques. Applied Technologies. Springer International Publishing, pp. 133–144, (2020). https://doi.org/10.1007/978-3-030-42520-3_11
9. Shieh, M.-D., Yang, C.-C.: Multiclass SVM-RFE for product form feature selection. Expert Syst. Appl. **35**(1–2), 531–541 (2008). https://doi.org/10.1016/j.eswa.2007.07.043
10. Curilem, G., Vergara, J., Fuentealba, G., Acuña, G., Chacón, M.: Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. J. Volcanol. Geoth. Res. **180**(1), 1–8 (2009). https://doi.org/10.1016/j.jvolgeores.2008.12.002
11. Avesani, R., Azzoni, A., Bicego, M., Orozco-Alzate, M.: Automatic classification of volcanic earthquakes in hmm-induced vector spaces, pp. 640–647 (2012). https://doi.org/10.1007/978-3-642-33275-3_79

12. Lara-Cueva, R., Carrera, E. V., Morejon, J. F., Benitez, D.: Comparative analysis of automated classifiers applied to volcano event identification. In: 2016 IEEE Colombian Conference on Communications and Computing (COLCOM), pp. 1–6. IEEE (2016). https://doi.org/10.1109/ColComCon.2016.7516377
13. Lara, F., León, R., Lara-Cueva, R., Tinoco-S, A.F., Ruiz, M.: Detection of volcanic microearthquakes based on homomorphic deconvolution and sta/lta. J. Volcanol. Geoth. Res. **421**, 107439 (2022). https://www.sciencedirect.com/science/article/pii/S0377027321002687
14. Lara, F., Lara-Cueva, R., Larco, J.C., Carrera, E.V., León, R.: A deep learning approach for automatic recognition of seismo-volcanic events at the cotopaxi volcano. J. Volcanol. Geoth. Res. **409**, 107142 (2021)
15. Canário, J.P., Mello, R., Curilem, M., Huenupan, F., Rios, R.: In-depth comparison of deep artificial neural network architectures on seismic events classification. J. Volcanol. Geoth. Res. **401**, 1–6 (2020). https://doi.org/10.1016/j.jvolgeores.2020.106881
16. Peng, P., He, Z., Wang, L., Jiang, Y.: Automatic classification of microseismic records in underground mining: a deep learning approach. IEEE Access **8**, 863–876 (2020)
17. Liu, L., Song, W., Zeng, C., Yang, X.: Microseismic event detection and classification based on convolutional neural network. J. Appl. Geophys. **192**, 104380 (2021)
18. López-Pérez, M., García, L., Benítez, C., Molina, R.: A contribution to deep learning approaches for automatic classification of volcano-seismic events: deep Gaussian processes. IEEE Trans. Geosci. Remote Sens. **59**(5), 3875–3890 (2021)
19. Pérez, N., Benítez, D., Grijalva, F., Lara-Cueva, R., Ruiz, M., Aguilar, J.: Eseismic: towards an ecuadorian volcano seismic repository. J. Volcanol. Geoth. Res.. 106855 (2020)
20. Galarza, C., Vega, J.: Generación de señales sintéticas de eventos sismo-volcánicos del volcán cotopaxi a través de un modelo de red neuronal adversario generativo condicional (2022)
21. Van Dyk, D.A., Meng, X.-L.: The art of data augmentation. J. Comput. Graph. Stat. **10**(1), 1–50 (2001)
22. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Text data augmentation for deep learning. J. Big Data **8**, 1–34 (2021)
23. Kenneth, B.: The seismic wavefield. Introduction and Theoretical Development, Volume I. Cambridge Univ. Press, USA (2001)
24. Chouet, B.A., Page, R.A., Stephens, C.D., Lahr, J.C., Power, J.A.: Precursory swarms of long-period events at redoubt volcano (1989–1990), Alaska: their origin and use as a forecasting tool. J. Volcanol. Geoth. Res. **62**(1–4), 95–135 (1994)
25. Rousselet, G.A.: Does filtering preclude us from studying ERP time-courses? Front. Psychol. **3**, 131 (2012)
26. Khan, I.D., Khan, M.H., Farooq, O., Khan, Y.U.: A comparative analysis of seizure detection via scalogram using googlenet, alexnet and squeezenet. In: 2021 Smart Technologies, Communication and Robotics (STCR), pp. 1–5. IEEE (2021)
27. Zhang, Z., et al.: Analyzing bulk flow characteristics of debris flows using their high frequency seismic signature. J. Geophys. Res. Solid Earth **126**, 12 (2021)
28. Lara-Cueva, R., Larco, J.C., Benitez, D.S., Perez, N., Grijalva, F., Ruiz, M.: On finding possible frequencies for recognizing microearthquakes at cotopaxi volcano: a machine learning based approach. J. Volcanol. Geoth. Res. **407**, 107092 (2020). https://doi.org/10.1016/j.jvolgeores.2020.107092
29. Sifuzzaman, M., Islam, M.R., Ali, M.Z.: Application of wavelet transform and its advantages compared to fourier transform (2009)
30. Ngui, W.K., Leong, M.S., Hee, L.M., Abdelrhman, A.M.: Wavelet analysis: mother wavelet selection methods. Appl. Mech. Mater. **393**, 953–958 (2013)
31. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). https://www.deeplearningbook.org

32. Yang, H.-F., Dillon, T.S., Chen, Y.-P.P.: Optimized structure of the traffic flow forecasting model with a deep learning approach. IEEE Trans. Neural Networks Learn. Syst. **28**, 2371–2381 (2017)
33. Zhou, P., Han, J., Cheng, G., Zhang, B.: Learning compact and discriminative stacked autoencoder for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **57**, 4823–4833 (2019)
34. Low, C.-Y., Park, J., Teoh, A.B.-J.: Stacking-based deep neural network: deep analytic network for pattern classification. IEEE Trans. Cybern. **50**, 5021–5034 (2020)
35. Zeledón-Córdoba, M., Sánchez-Solís, J., Coto-Jiménez, M.: Reconstructing fundamental frequency from noisy speech using initialized autoencoders. IEEE Lat. Am. Trans. **18**(10), 1724–1731 (2020)
36. Poier, P.P., Jaffrelot Inizan, T., Adjoua, O., Lagardere, L., Piquemal, J.-P.: Accurate deep learning-aided density-free strategy for many-body dispersion-corrected density functional theory. J. Phys. Chem. Lett. **13**(19), 4381–4388 (2022)
37. Lara-Cueva, R., Larco, J.C., Benitez, D.S., Perez, N., Grijalva, F., Ruiz, M.: On finding possible frequencies for recognizing microearthquakes at cotopaxi volcano: a machine learning based approach. J. Volcanol. Geoth. Res. **407**, 107092 (2020)
38. Lara-Cueva, R.A., Benítez, D.S., Carrera, E.V., Ruiz, M., Rojo-Álvarez, J.L.: Automatic recognition of long period events from volcano tectonic earthquakes at cotopaxi volcano. IEEE Trans. Geosci. Remote Sens. **54**(9), 5247–5257 (2016)
39. Pérez, N., Venegas, P., Benitez, D., Grijalva, F., Lara, R., Ruiz, M.: Benchmarking seismic-based feature groups to classify the cotopaxi volcanic activity. IEEE Geosci. Remote Sens. Lett. 1–5 (2020)
40. Titos, M., Bueno, A., García, L., Benítez, C., Segura, J.C.: Classification of isolated volcano-seismic events based on inductive transfer learning. IEEE Geosci. Remote Sens. Lett. **17**(5), 869–873 (2019)

Implementing IoT Technology for Energy Optimization in Rooftop Solar PV Plants: A Concentrated Validation Study

Fabian Alonso Lara-Vargas[1,2](✉), Jimena de la Ossa-Rivera[1], Santiago Jaramillo-Mira[1], Carlos Vargas-Salgado[2], Jesus Aguila-Leon[3], and Evelyn Villabon-Lopez[4]

[1] Universidad Pontificia Bolivariana, Monteria, Colombia
fabian.lara@upb.edu.co
[2] Universitat Politécnica de Valéncia, Valencia, Spain
[3] Universidad de Guadalajara, Jalisco, Mexico
[4] URRA S.A E.S.P, Monteria, Colombia

Abstract. The present study presents a low-cost Internet of Things (IoT) monitoring system for rooftop solar photovoltaic systems to promote sustainable energy solutions. Using a Raspberry Pi Pico W and low-cost sensors, the system enables real-time monitoring of important variables such as solar panel temperature, solar radiation, inverter temperature, and indoor temperature. A case study conducted in Montería, Colombia, validated the accuracy of the proposed system by comparing its measurements with those of commercial meters using metrics such as Root Mean Square Error (RMSE), correlation, Mean Absolute Error (MAE), and R - Square (R^2). The system achieved comparable correlation values between 0.84 and 0.98, RMSE between 74.6 and 1.2, MAE between 3.22 and 1.2, and R^2 between 0.88 and 0.98. The results demonstrate the system's ability to increase the performance of solar systems and thus support a sustainable energy transition. This study highlights the need for efficient monitoring and management of solar photovoltaic systems to maximize potential and minimize environmental impact. By examining precise and reliable monitoring methods, electronic prototypes for photovoltaic monitoring systems will be validated. These results highlight the importance of integrating IoT technology and data analytics to improve the automation, control, and monitoring of photovoltaic systems and promote renewable energy adoption and a sustainable future.

Keywords: Solar PV plant · IoT monitoring · Energy sustainability · Inverter

1 Introduction

Solar energy is a crucial part of the development of renewable energy sources worldwide. The Earth receives a remarkable amount of energy from the sun, far exceeding its annual consumption. Therefore, solar energy is widely regarded as a practical solution to meet global energy needs [1]. The growth and sustainability of solar energy as a

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 14–26, 2025.
https://doi.org/10.1007/978-3-031-88854-0_2

renewable energy source is evident in the shift of public subsidies toward profitability in the photovoltaic industry in the 2010s. This shift shows the significant development and environmental friendliness of solar energy [1]. Although forecasts suggest that renewable energy sources could potentially meet half of the world's energy needs by 2040, largely thanks to solar power, only 14% of energy needs are currently met by renewables [2]. It is important to highlight the importance of continued advancements and funding to realize the full potential of solar PV technology.

Monitoring photovoltaic solar systems is critical to achieving optimal performance and efficiency while minimizing their negative environmental impact. It is essential to optimize the potential of these plants [3]. Maintaining the optimal performance of solar systems is crucial, especially to reduce the environmental footprint. Efficient performance requires regular cleaning and prediction using machine learning algorithms [4].

Photovoltaic systems, a popular renewable energy source, can experience fluctuations in performance due to environmental factors such as temperature, cloud cover, and altitude [4]. Evaluating performance considering environmental factors is crucial to obtaining accurate and reliable efficiency ratings for solar photovoltaic systems, as demonstrated in previous studies [5]. Monitoring photovoltaic solar systems is crucial for optimizing their performance and reducing their environmental impact. Machine learning techniques can be used to predict performance and improve operations [4], analyze environmental factors [5], solar tracking, intelligent monitoring, dust impact studies, and the cost-effectiveness of remote monitoring [5], and contribute to the efficient management of solar photovoltaic plants.

Several studies have examined the use of sensors and data analysis to monitor photovoltaic systems; However, precise and reliable methods are still required. The literature emphasizes the importance of sensors and data analysis in monitoring photovoltaic systems [6]. These methods are essential for optimizing the performance and ensuring the reliability of solar energy systems. Integrating advanced sensors with wireless communication protocols can potentially improve the resolution of surveillance systems and enable near real-time data collection [6]. Internet of Things (IoT) technology, which incorporates various sensors, is crucial for improving the automation, control, and monitoring of photovoltaic systems [7]. However, there are challenges to be addressed, such as economic viability, data management, and security concerns [7]. IoT-based photovoltaic (PV) system monitoring systems offer significant benefits in optimizing performance, reducing operational costs, and improving reliability. These systems enable real-time data collection and analysis of various parameters such as voltage, current, temperature, and irradiance at both panel and system levels [8]. The integration of IoT technology enables remote monitoring and control, making it particularly useful for PV systems in remote locations [9].

Implementing IoT-based monitoring systems not only enhances the efficiency of PV systems but also results in significant operation and maintenance (O&M) cost savings. A study conducted in Kenya showed that annual O&M costs could be reduced by 47–95% compared to locally monitored systems, with site-specific cost savings ranging from $2,040 to $3,096 per year [10]. This underscores the economic benefits of introducing IoT-based monitoring solutions for PV systems, making a compelling case for their practical advantages. IoT platforms such as Ubidots and open-source tools such as Arduino are used to monitor solar panel performance, detect problems, and provide real-time data on electricity production. This enables preventive maintenance and efficient management of solar power systems [11]. Advanced monitoring systems such as Fusion Solar and iSolarCloud collect and analyze real-time data on power generation and system performance, enabling timely maintenance and maximizing power production [12]. However, their operating cost is considerable.

In addition, thermal management strategies can significantly increase the efficiency of photovoltaic cells because the temperature of the solar panel plays a crucial role in the efficiency of energy production [4]. Environmental factors such as temperature and light influence the performance of photovoltaics and reduce their efficiency [5]. Photovoltaic systems with installation angles below 20° have a minimal impact on energy production, as various studies show [13]. The quality of the output power of PV systems can be influenced by various factors, including an increase in the internal temperature of the inverter. High-performance management without appropriate temperature control can lead to temperature-related performance reduction. However, it is necessary to develop optimal operating modes to mitigate these problems [6].

Numerous studies have focused on the development and validation of models and systems for tracking and predicting the performance of solar photovoltaic systems. For example, [14] examined the use of machine learning algorithms to detect faults in photovoltaic panels and indicated that research is currently underway on intelligent monitoring systems. Likewise, [6] proposed a novel real-time monitoring system with higher resolution and integrated solar tracking, suggesting advances in monitoring technologies. Nevertheless, it is important to recognize that while these studies contribute to the field, they may not cover all aspects of electronic prototype validation. Studies such as [15] and [16] presented methodologies for validating photovoltaic solar plant models through dynamic simulation of hybrid data, which implies that the techniques for validation are effectively the subject of research, although these studies focus more on the accuracy of the model in relation to its integration into the electrical grid rather than on the electronic prototypes themselves. Although research has been conducted on solar photovoltaic system monitoring and model validation systems, there is a lack of focus on the precision and reliability validation of electronic prototypes.

The study system was installed on the roof of a building in the city of Monteria, Colombia, at coordinates 8.803750 and −75.850125. The connection to the power grid is via a substation on the roof of the building. The system consisted of 240 panels with an output of 400 W and achieved a peak output of 96 kW. The CC to CA conversion system includes two inverters, each 36 kW, with a total output of 72 kW.

This research addresses this lack of understanding by evaluating the accuracy and dependability of an electronic device designed for monitoring a rooftop solar photovoltaic system. The key question is: How accurate and reliable is this electronic device compared to the measurements of a commercial weather station, a temperature data logger, and the inverter's internal measurement systems when inverters are installed on the roof? The system includes a comprehensive network of sensors that collect and transmit measurement data every hour. The external sensors measure the temperature of the solar modules in °C and the solar radiation in W/m^2 at the same angle as the solar modules, while the internal sensors measure the indoor temperature in °C and the temperature of the inverter in °C.

This study proposes a cost-effective and innovative strategy for monitoring photovoltaic systems by integrating low-cost hardware and sensors into an IoT network. This approach addresses gaps in affordability, real-time monitoring and thermal management of critical components such as inverters. The new strategy not only makes advanced monitoring of photovoltaic systems accessible to a larger number of users but also establishes a new standard for optimizing renewable energy management. Furthermore, this strategy can potentially inspire future research and development in this sector. The aim of this study is to improve the performance of photovoltaic system monitoring and reduce costs by developing a low-cost, IoT-integrated monitoring system that promotes a more sustainable energy transition.

The prototype was validated by comparing its data over a two-month period with that of a commercial weather station, a temperature data logger, and internal measurement systems of the solar inverter. Statistical analysis included calculation of RMSE, MAE, Spearman correlation, and R^2 values and comparison of variables. Hourly data was stored in the cloud and accessible via a website. Creating an IoT system involves eight separate steps: identifying the variables to be measured, designing the design of the integrated system, creating a prototype, and validating methods and testing periods. Steps five and six focus on selecting and collecting data from commercial devices, while step seven analyzes the results and step eight presents the conclusions.

2 Methodology

This section outlines the methodology employed to create and test an electronic prototype for monitoring a solar photovoltaic plant. This IoT-based system was designed to gather data on environmental and operational variables with the goal of contributing to future research and advancements in the field of solar energy. The methodology is divided into

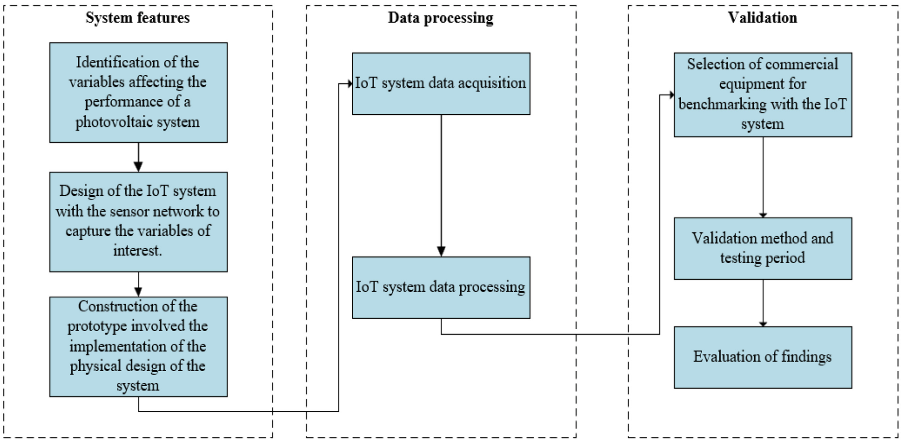


Fig. 1. Methodology workflow.

eight sections, detailing the process from variable identification to drawing conclusions. See Fig. 1.

2.1 Identification of Variables

The main factors affecting PV system performance were identified through a literature review and included solar radiation, module temperature, inverter temperature, and environmental conditions around the inverter. The energy production of PV modules is directly influenced by solar radiation, with higher radiation leading to increased electricity production [17]. However, the temperature of a photovoltaic module is a critical parameter, as a higher module temperature can significantly reduce the efficiency and output power of the system [17]. Solar radiation and module temperature have a direct effect on photovoltaic performance. However, the temperature of the inverter and the ambient conditions are also crucial. Extreme ambient temperatures can negatively impact the performance of solar panels by increasing their operating temperature [18]. Inverters that convert direct current to alternating current can exhibit reduced efficiency and reliability at high temperatures and humidity [18]. Efficiently managing the complex interactions between solar radiation, photovoltaic module temperature, inverter temperature, and inverter environmental conditions is critical to optimizing the performance and longevity of solar energy systems [19].

The study identified the key variables to be observed in the case study, which featured a rooftop photovoltaic system with inverters installed inside the roof and a temperature control system within. The installation of rooftop equipment, including inverters (1) and photovoltaic plants (2), is illustrated in Fig. 2 the discussion is presented in Sect. 2.2.

Fig. 2. Rooftop solar photovoltaic installation for the case study.

2.2 Integrated IoT System Design

The IoT system, designed for real-time data transmission, focuses on sensor selection, Raspberry Pi Pico W microcontroller integration, and a user interface for visualization.

The architecture and location are shown in Fig. 3. The sensors chosen were low-cost, compatible with the Raspberry Pi system, and provided adequate precision for their price.

- Davis 6450 Pyranometer: This instrument has a spectral measurement range of 400–1100 nm and a resolution of 0.1, with a corresponding range of units [20].
- The Max6675-based Type J thermocouple in this prototype serves to obtain the panel temperature data, with a temperature resolution of 0.25 °C and a temperature range of −200 °C to 1300 °C [21].
- For data processing, a Raspberry Pi Pico W microcontroller was utilized owing to its capabilities for Internet of Things (IoT) projects [22].
- The infrared sensor, MLX90614, measured the temperature within the inverter. This sensor has a range of −70 °C to +380 °C and a precision of ±0.5 °C [23].
- DHT11 temperature and humidity sensors. The temperature of the sensor ranged from 0 °C to 50 °C [24].

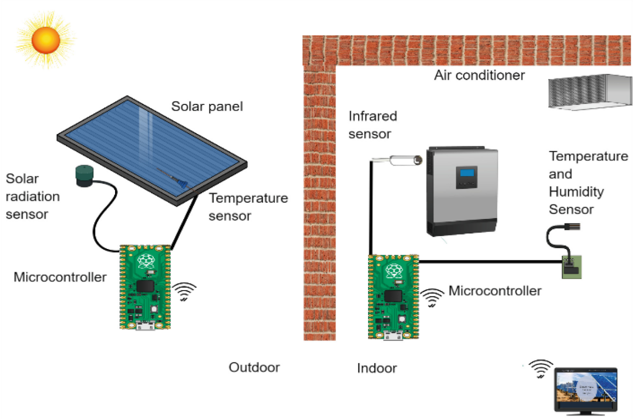


Fig. 3. Architecture of the IoT system and sensor network for monitoring variables.

2.3 Prototype Construction

The construction of the prototype involved the implementation of the physical design of the system. Connections were made between the selected sensors and the microcontroller, the data capture and transmission logic were programmed using Micro Python, and the Wi-Fi connection was established for data transmission to the cloud. However, a control algorithm must be developed. In addition, the development of algorithms and programming for the controls is required, as shown in the flowchart in Fig. 4.

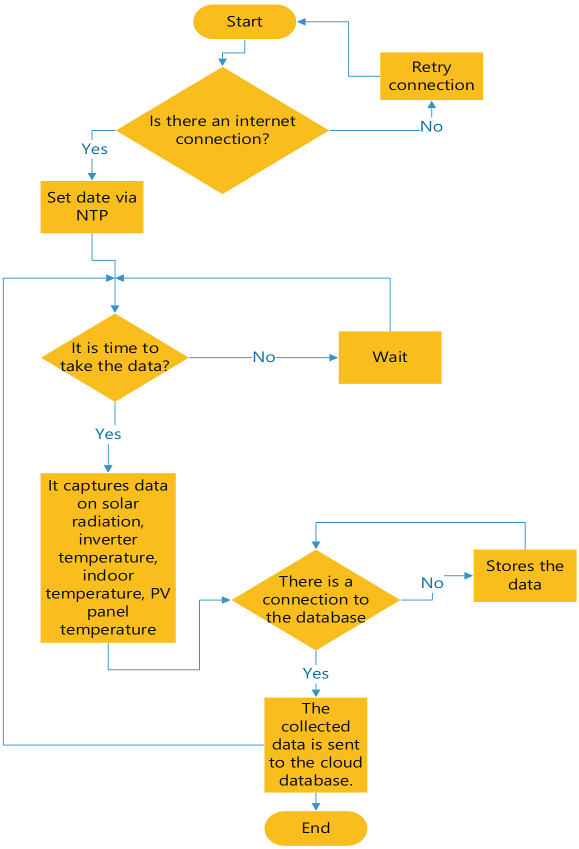


Fig. 4. Flow chart of microcontroller programming.

2.4 Data Acquisition

The data were acquired between November 1 and December 31, 2023, at 60-min intervals and included measurements of solar radiation in watts per square meter (W/m^2), indoor temperature, and solar module and inverter temperature in degrees Celsius (ºC).

2.5 Data Analysis

The data processing process was automated through the prototype, collecting measurements every hour, and to a MySQL database hosted by GoDaddy. An interface was developed to visualize and analyze the data in real-time. The visual interface can be admired in Fig. 5.

Fig. 5. Visual interface of the website with cloud storage.

2.6 Selection of Commercial Equipment for Benchmarking

For the comparative evaluation, an Ambient Weather W2000 weather station with a resolution of 0.001 W/m^2 and 0.1 °F, 0.1%RH and the Elitech Gsp-6 datalogger with a resolution of 0.1 °C/°F, 0.1%RH were selected as reference equipment to validate the measurements of the prototype, given its proven accuracy and reliability and affordable cost. Figure 6 shows the AW W2000 weather station (1), the IoT system (2) that collects signals from the solar panel temperature sensor and the solar radiation sensor (3).

Fig. 6. Configuration of the IoT system and reference weather station for data collection

2.7 Validation Methods and Testing Period

The study utilized statistical methods, such as correlation, root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2), to evaluate the data collected by the prototype in comparison to the information obtained by commercial teams. The testing period was determined to be two months, from November to December 2023, based on data availability during that time frame. RMSE calculates the error magnitude by dividing the squared differences by N and taking the square root, indicating model precision with a lower score [25]. This metric measures the size of the errors between the values predicted by the model V*predicted* and the actual values V*target*. Calculate a sum of squares of the difference of these values, divide it by the sample number N and take a root from the result of this operation. See Eq. 1.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n=1}^{N} \left(\mathrm{V}_{\mathrm{predicted}} - \mathrm{V}_{\mathrm{target}}\right)^2}{\mathrm{N}}} \tag{1}$$

The method of correlating variables is an essential statistical tool with wide-ranging applications across all disciplines. Its use is vital for understanding the relationships between variables, despite potential pitfalls in its application and interpretation [26]. The Mean Absolute Error (MAE) is defined as the average of the absolute differences between the predicted and actual values, without considering the direction of those errors [26]. See Eq. 2.

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{k}=1}^{\mathrm{n}} \left|\mathrm{yi} - \hat{\mathrm{y}}\mathrm{i}\right| \tag{2}$$

The number of observations is denoted by "n", "yi" represents the actual value and "ŷi" denotes the predicted value. R-squared (R^2) measures the accuracy of a model's predictions of quantitative variables. It was calculated as the average of the absolute errors between the predicted values and actual observations [26]. This metric assesses the proportion of the variance in the measurements that is predictable from the independent variables in the model; It is calculated as 1 minus the sum of the squares of the differences between the actual measured values yi, actual, and the predicted values yi, predicted and divided by the sum of the squares of the difference between the actual measured values and the average of these values ȳ, actual. Where N denotes the number of samples in the calibration and validation sets. See Eq. 3.

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{yi, actual} - \ \mathrm{yi, predicted})^2}{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{yi, actual} - \bar{\mathrm{y}}\mathrm{, actual})^2} \tag{3}$$

2.8 Evaluation of Findings

The conclusions drawn from the project highlight the effectiveness of the prototype in providing precise and reliable measurements of key variables in the monitoring of solar photovoltaic plants. These findings validate the prototype's potential as a tool for future research in the field of solar energy, emphasizing areas for improvement and recommendations for its implementation in other contexts or project scales. The following section describes the process used to analyze the results.

3 Results

In conducting the analysis of correlation, root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) for the months of November and December 2023, the findings were as outlined in Table 1.

Table 1. Comparative metric results of the pattern measure and IoT systems.

	RMSE	Correlation	MAE	R^2
Solar Radiation	74,63	0,98	1,63	0,88
Inverter Temperature	3,22	0,84	3,22	0,69
Indoor Temperature	1,20	0,94	1,20	0,74
Module Temperature	1,23	0,99	1,23	0,98

Based on these assertions, it can be stated that the prototype exhibits a good level of confidence in its correlation with the data of the reference element with regard to the metrics used, highlighting the measurement of the temperature of the solar module and the temperature within the installation.

The IoT system demonstrated outstanding performance in solar radiation capture, with an RMSE of 74.63 W/m^2 and a correlation of 0.98, confirming its high accuracy in real-time monitoring. The slight discrepancy in RMSE can be attributed to environmental factors such as brief cloud cover or small sensor mismatches. In addition, the solar radiation sensor of the weather station is installed vertically, while the radiation sensor of the IoT system is installed in the same plane as the solar panel, i.e. at 8° from horizontal.

The PV module temperature showed strong agreement with reference measurements, with an R^2 value of 0.98 and an RMSE of only 1.23 °C. Accurate temperature measurement is essential as it directly impacts the efficiency and energy production of solar panels. These results highlight the ability of the IoT system to provide reliable data for thermal management, thereby optimizing the performance of PV modules.

For inverter temperature, the system showed a lower correlation (0.84) and an RMSE of 3.22 °C, suggesting the presence of additional factors, such as inverter load or ventilation conditions, that are not taken into account in the current model. These results suggest that the model could benefit from further adjustments to improve the accuracy of the inverter temperature prediction or to analyze a different position of the temperature sensor.

Indoor temperature measurements were also accurate, with an RMSE of 1.20 °C and a correlation of 0.94. Figure 7 describes the relationship between the variables of the patterns and the IoT system.

The analysis of the graphs indicates that the estimation methods for solar radiation, panel temperature, inverter temperature, and interior temperature are accurate and well correlated with actual measurements, with close medians and quartiles and a comparable spread of data. However, the sensor mean of the IoT solar radiation sensor is a little out of line with the mean of the reference system (solar radiation sensor of the weather station).

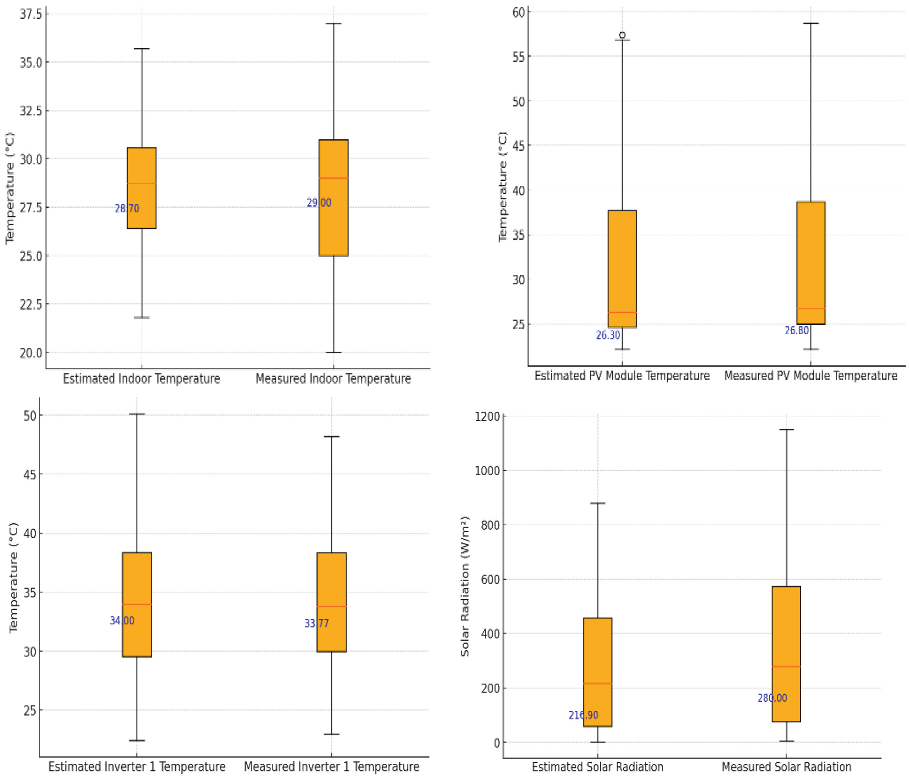


Fig. 7. A comparative diagram of the variables measured in IoT patterns and the estimated system variables.

On the other hand, the cost of the IoT system was $240 compared to $450 for commercial devices, although each device displays data independently, unlike the IoT system that integrates all data into a single database and cloud-based visualization system.

Future research could produce improvements from several perspectives. From a methodological perspective, extending the validation period, including additional variables, and optimizing sensor calibration will improve the accuracy and reliability of the IoT monitoring system. From the perspective of results, integrating advanced statistical techniques and conducting sensitivity and error analysis will deepen the understanding of system performance.

4 Conclusion

This paper presents an analysis of the design, launch, and validation of an economic prototype system for monitoring solar photovoltaic installations through the Internet of Things (IoT). An innovative approach that integrates accessible hardware, such as the Raspberry Pi Pico W, and a selection of specific sensors. This study addresses crucial challenges in monitoring photovoltaic plants, including affordability, real-time monitoring accuracy, and efficient temperature management of the critical components. The quantitative results obtained during the prototype validation highlight its precision and reliability compared to those with commercial reference equipment. These results underscore the effectiveness of the prototype in capturing critical variables with precision for optimal operation and maintenance of photovoltaic plants.

The research was limited to a two-month validation period, which may not fully capture seasonal variability and its impact on the system. The selection of variables focused on specific parameters, excluding other potentially relevant factors for the performance of photovoltaic facilities, such as wind speed and dust accumulation on the panels. This research not only establishes a new cost-effective monitoring system but also paves the way for wider adoption of such technology in photovoltaic power plant management. In the future, protocols for calibration and adjustment can be developed to improve accuracy and better align values with patterns.

References

1. Mitrašinović, A.M.: Photovoltaics advancements for transition from renewable to clean energy. Energy **237** (2021). https://doi.org/10.1016/j.energy.2021.121510
2. Fan, X., Liu, W., Zhu, G.: Scientific linkage and technological innovation capabilities: international comparisons of patenting in the solar energy industry. Scientometrics **111**(1) (2017). https://doi.org/10.1007/s11192-017-2274-5
3. Fabian, M., Vargas, A.L., Ángel, M.M., Padilla, O., Carlos, P., Salgado, V.: Comparative experimental analysis of the annual energy production of a 72kWn photovoltaic solar power plant installed on a roof for self-consumption in the city of Monteria using PVsyst, PVGIS and SAM **1**(43), 51–56 (2024). https://doi.org/10.24054/rcta.v1i43.2807
4. Ra, N., Varman, S., Antony Joseph, K., Bhattacharjee, A.: Prediction of optical performance of solar PV under the impact of natural dust accumulation: machine learning approach. In: 2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies, GlobConHT 2023 (2023). https://doi.org/10.1109/GlobConHT56829.2023.10087364
5. Wang, Z., Li, Y., Wang, K., Huang, Z.: Environment-adjusted operational performance evaluation of solar photovoltaic power plants: a three stage efficiency analysis (2017). https://doi.org/10.1016/j.rser.2017.03.119
6. Sadineni, S.B., Realmuto, J.D., Boehm, R.F.: An integrated performance monitoring and solar tracking system for utility scale PV plants. Am. Soc. Mech. Eng. Power Div. POWER (2011). https://doi.org/10.1115/POWER2011-55243
7. Rajeshwar Reddy, K., Arabelli, R.R., Rajababu, D., Mahender, K.: Solar power generation system with IOT based monitoring and controlling using different sensors and protection devices to continuous power supply. In: IOP Conference Series: Materials Science and Engineering (2020). https://doi.org/10.1088/1757-899X/981/3/032017
8. Rouibah, N., Barazane, L., Benghanem, M., Mellit, A.: IoT-based low-cost prototype for online monitoring of maximum output power of domestic photovoltaic systems. ETRI J. **43**(3) (2021). https://doi.org/10.4218/etrij.2019-0537
9. Khan, M.S., Haque, A., Bharath, K.V.S.: Real-time solar inverter parameter monitoring for photovoltaic systems. In: 2019 International Conference on Power Electronics, Control and Automation, ICPECA 2019 - Proceedings (2019). https://doi.org/10.1109/ICPECA47973.2019.8975519

10. Njoka, F., Thimo, L., Agarwal, A.: Evaluation of IoT-based remote monitoring systems for stand-alone solar PV installations in Kenya. J. Reliab. Intell. Environ. **9**(3) (2023). https://doi.org/10.1007/s40860-022-00190-5
11. Wanjari, A., et al.: Hybrid renewable energy generator with IOT monitoring **1**(1) (2023). https://doi.org/10.46610/jaeed.2023.v01i01.005
12. Thi Ninh, T.K.: Utilizing monitoring systems for information management in rooftop solar photovoltaic plants. In: 2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD), pp. 176–180 (2023). https://doi.org/10.1109/BCD57833.2023.10466324
13. Alici, H., Esenboga, B., Oktem, I., Demirdelen, T., Tumay, M.: Designing and performance analysis of solar tracker system: a case study of Çukurova region. Des. Anal. Appl. Renew. Energy Syst. (2021). https://doi.org/10.1016/B978-0-12-824555-2.00004-6
14. Yun, L., Bofeng, Y., Dan, Q., Fengshuo, L.: Research on fault diagnosis of photovoltaic array based on random forest algorithm: In: Proceedings of 2021 IEEE International Conference on Power Electronics, Computer Applications, ICPECA 2021 (2021). https://doi.org/10.1109/ICPECA51329.2021.9362559
15. Zhao, D., Hu, D., He, J., Zhang, L., Chen, N.: Model validation of solar PV plant with hybrid data dynamic simulation based on fast-responding generator method. In: MATEC Web of Conferences (2016). https://doi.org/10.1051/matecconf/20166502006
16. Zhao, D.W., et al.: Hybrid data simulation based model validation method for solar PV plant. In: IET Conference Publications (2015).https://doi.org/10.1049/cp.2015.0533
17. Gupta, V., Sharma, M., Pachauri, R., Babu, K.N.D.: Impact of hailstorm on the performance of PV module: a review (2022). https://doi.org/10.1080/15567036.2019.1648597
18. Atyia, T., Qasim, M.: Evaluating the impact of weather conditions on the effectiveness and performance of PV solar systems and inverters. NTU J. Renew. Energy **5**(1) (2023). https://doi.org/10.56286/ntujre.v5i1.551
19. Kulkarni, V.V., Kulkarni, V.A.: Performance optimization of photovoltaic systems using thermoelectric cooling system. In: 2022 International Conference on Futuristic Technologies, INCOFT 2022 (2022). https://doi.org/10.1109/INCOFT55651.2022.10094413
20. Vantage Pro2TM Accessories, Davis: 6450 Pyranomètre MANUAL EN DAVIS (2014). www.davisnet.com
21. Maxim Integrated Products: MAX6675 user manual (2021). www.maximintegrated.com
22. Raspberry Pi Ltd.: User manual raspberry pi Pico W (2023)
23. P. By ALLDATASHEETCOM: MLX90614 family single and dual zone infra red thermometer in TO-39 features and benefits.
24. Asair: Digital temperature & humidity module DHT11 user manual (2019). www.aosong.com
25. Hodson, T.O.: Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not (2022). https://doi.org/10.5194/gmd-15-5481-2022
26. Cho, D.S., Chung, C., Kim, J., Ahn, S., Park, S., Park, H.S.: Analysis on reports of statistical testings for correlation and regression. Korean J. Women Health Nurs. **14**(3) (2008). https://doi.org/10.4069/kjwhn.2008.14.3.213

Performance Testing of Large Language Models in the Context of Industry 4.0

Pedro José González Díaz(✉), Ailín Orjuela Duarte, William Mauricio Rojas Contreras, and Luz Marina Santos Jaimes

Universidad de Pamplona, Km. 1 Vía Bucaramanga, Pamplona, Colombia
{pedro.gonzalez,aorjuela,mrojas,lsantos}@unipamplona.edu.co
http://www.unipamplona.edu.co

Abstract. Large Language Models (LLMs) are a type of artificial intelligence capable of processing and generating natural language. These models are trained on vast amounts of data, such as text and code, which enables them to perform various tasks like text generation, language translation, question answering, text summarization, etc.. The purpose of this research was to find an LLM that meets the following requirements: 1)easy to implement with an understanding of the Spanish and English languages, and 2)accurately answers diagnostic questions and action plans related to Industry 4.0 (I4.0). Some open-source LLMs were selected, each with 7B parameters and quantizations of Q2 and Q4 for the application of a first set of general tests. The results showed that the model with the best performance in both languages English and Spanish was Mistral. A second set of Spanish tests was then conducted with and without Retrieval-Augmented Generation (RAG), using three documents related to the topic of I4.0. The results demonstrated Mistral's capability with RAG to answer questions on the studied context with 95 % accuracy.

Keywords: Large Language Model · Industry 4.0 · Retrieval-Augmented Generation · Quantization

1 Introduction

The Fourth Industrial Revolution (Industry 4.0 - I4.0) marks a new productive era focused on automation, digitalization, and real-time connectivity of production systems [14]. The recent emergence of Large Language Models (LLMs) demonstrates the potential of Artificial Intelligence (AI), revealing new opportunities in I4.0 and smart manufacturing. However, there is a significant gap in the application of these LLMs in the industry, primarily due to their training in general knowledge rather than in specific expertise. This specialized domain knowledge is vital to effectively address the complex needs of industrial applications [16].

The novelty of this work lies in the ability of the LLM to provide an accurate diagnosis of the level of I4.0 adoption based on a survey conducted to companies,

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 27–40, 2025.
https://doi.org/10.1007/978-3-031-88854-0_3

and to identify the steps to follow for implementing I4.0 technologies. This work presents the results of the application of two sets of local tests to some selected LLMs, which were chosen at the time of the project for being free, having low computational requirements, being supported by large companies, and handling 7 billion parameters.

An evaluation was conducted using a quantitative response-to-question metric methodology. The first set of tests was performed with two different quantizations for each model: 4 bits and 2 bits. The desktop program LM Studio was used, which allows downloading different LLMs locally, as long as the computer meets the minimum recommended hardware requirements. The LLMs can be downloaded with different quantizations and a direct connection to explore more about the model to be tested using the Huggingface library.

The responses generated by the models were evaluated in aspects such as logical reasoning, natural language understanding, code comprehension, code generation, understanding of technical concepts, and human-machine interaction. For the second set of tests, the desktop application GPT4ALL was used, which allows running the model with Retrieval-Augmented Generation (RAG) locally. The test consisted of five new questions focused on I4.0 technologies and their implementation, questions that were related to the content of three documents provided with RAG. The introduction is presented in this section. The rest of the article presents related work in Sect. 2, methodology in Sect. 3, results in Sect. 4, and conclusions in Sect. 5.

2 Related Work

Related works on the evaluation of LLMs in general and the application of LLMs to Industry 4.0 were identified.

2.1 Models Evaluation

In [11], a panorama of the evaluation criteria for LLMs is given: accuracy, ethics, fairness, generalization, robustness, and reasoning. Similarly, the work classifies the evaluation metrics into three types: multiple classification metrics (MC), token similarity metric (TS), and question-answering metrics (QA).

In [21], LLMs GPT-4 and BERT were employed to evaluate responses to queries about "symptoms of diabetes" and "causes of migraines". A custom evaluation set was developed to measure responses based on three main criteria: factual accuracy, contextual relevance, and ethical robustness. Code generation through LLMs is one of the most widely used applications, following the popularity of tools like ChatGPT or GitHub Copilot. The objective in [8] was to conduct a study to assess the quality of code generated using different methodologies. In [23], 11 open-source LLMs and 4 closed-source LLMs were tested, and deficiencies in response security were found in some models between 20% and 50%. A system prompt emphasizing security substantially reduces the occurrence of insecure responses but does not completely prevent them. In [22], the capabilities of LLMs ChatGPT-4, Google Gemini, and Mistral 8×7B in text generation processes in English and Japanese were explored. Mistral showed balanced performance in both languages, and the study concludes that more diverse and inclusive linguistic data training leads to better results.

2.2 Application of LLMs to Industry 4.0

In [14], the main needs of ChatGPT for I4.0 are studied, identifying and reporting several associated characteristics, traits, and versatile capabilities of ChatGPT for I4.0, as well as important applications. In [16], an industrial large knowledge model (ILKM) is proposed, emphasizing its potential to revolutionize the smart manufacturing industry and comparing it with LLMs from eight perspectives. Finally, the 6S Principle is proposed as a guide for developing an ILKM in smart manufacturing. In [17], they present QUA4I (Question Answering for the I4.0), an intelligent virtual assistant for the I4.0 domain, including modules for automatic voice recognition and text-to-speech, and a custom demonstration designed to answer questions and extract information about the industry.

In [10], large models are categorized into three levels: large language models size (LLM), basic models for specialized areas and refined models for specific tasks. Its application in the oil and gas industry is still at an early stage and requires a large amount of computation; it points out that the quality and amount of data are not enough for training a large model because the data comes from underground environments. In [6], the authors present time series processing techniques using a RAG approach and the use of LLMs results in a model of large language that generates recovery-based fault diagnosis reports of historical data.

3 Methodology

This section provides detailed information about the AI-based language models, RAG, and LLMs selected for the study, as well as the set of tests conducted along with their metrics. RAG is an approach that combines text generation with information retrieval about Industry 4.0, instead of relying solely on a parametric language model. The tests were designed to measure the accuracy percentage of each of the models.

3.1 Large Language Models (LLM)

An LLM is an AI program that can recognize and generate text, among other tasks. LLMs are based on machine learning, specifically on a type of neural network called a transformer model, which helps understand how characters, words, and phrases function together [3]. An LLM is a program that has been provided with enough examples to recognize and interpret natural language or other types of complex data. LLMs are trained with massive datasets, many of which are collected from the Internet, utilizing thousands or millions of gigabytes of text.

Quantization (Q) is a technique to reduce computational and memory costs for inference, consisting of representing activation weights with low-precision data types, such as 8-bit integers, instead of the usual 32-bit floating-point type [12], resulting in an acceptable reduction in model capabilities and accuracy.

3.2 Retrieval-Augmented Generation (RAG)

LLMs may not include specific or updated information regarding the particular requirements of an organization, leading to incorrect responses that make the tool less useful. As a solution, RAG appears, optimizing the results of an LLM with contextualized and updated information specific to a company without needing to modify the underlying model and providing accurate responses to queries [26]. A traditional RAG system consists of an LLM, a vector database (to properly store external data), and a series of commands or questions. The process is divided into two parts: information input and enhanced information retrieval (see Fig. 1).

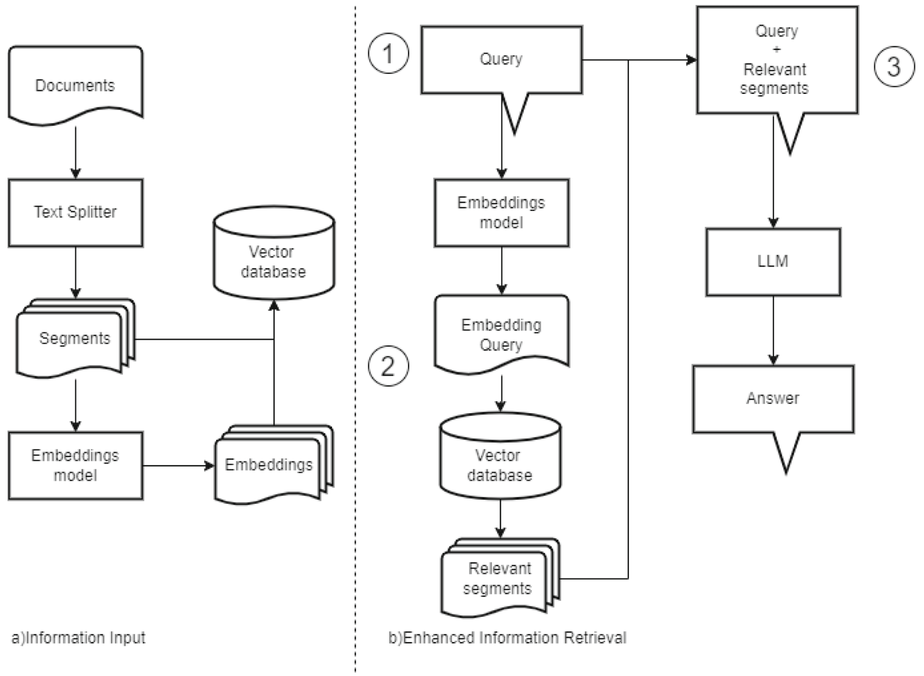


Fig. 1. Process in RAG

In the information input stage: the textual content of a document is extracted, divided into segments, and an embedding model is applied to produce the numerical representation of the text in vector form. The textual segments and their vector representation are stored in a repository.

In enhanced information retrieval: 1) The user makes a textual query or "prompt". 2) RAG enriches the prompt with data and facts obtained from an external database that contains information related to the user's question. 3) RAG sends the enriched prompt to the model, which subsequently generates a natural language response by leveraging all the knowledge it learned from the language provided by humans.

3.3 Fine-Tuning

Fine-tuning consists of refining a pre-trained base model to improve its specific task capabilities. This is easier and more economical than training a new model from scratch, especially in the case of deep learning models with millions or billions of parameters. By leveraging the model's prior training, the amount of processing and labeled data required to obtain large models adapted to specific use cases and business needs is reduced [13].

3.4 LLMs Under Study

Accessing the fine-tuning capabilities of GPT-4 requires a paid subscription; otherwise, the open-source domain is full of alternatives that offer a more accessible path to harness the power of LLMs. These open-source options democratize access to advanced AI technology, fostering innovation and inclusion in the rapidly evolving AI landscape. The development of language models like Gemma, Mistral, and Llama 2 represents a significant leap, with each model—developed by tech giants such as Google DeepMind, OpenAI, and Meta (Facebook AI), respectively—offering unique innovations and capabilities [15]. Additionally, at the time of the project, they were selected for being free, with low computational requirements, and backed with documentation and support.

Gemma: This model can be customized with fine-tuning techniques to excel in tasks important to users. It is available to run on applications and hardware, mobile devices, and hosted services. Gemma has been primarily designed for natural language processing in English and has two different models: Gemma 2B and Gemma 7B, with "B" referring to "billions of parameters" [9].

Mistral: This model stands out for focusing on creating AI models that are not only innovative but also open and accessible. The company's commitment to open-source software and its pioneering work in language models represent a significant shift in the AI landscape. By releasing many of its models and deployment tools under permissive licenses, Mistral contributes to the AI field while advocating for a paradigm shift [1].

Llama2: The model is the result of a collaboration between Meta, formerly Facebook, and Microsoft. Llama2 offers its code to any company, developer, or individual. Through the use of natural language processing algorithms, it can effectively understand and respond to user queries and requests, allowing for a much more natural and intuitive user experience. Llama 2 is a collection of three language models: 7B, 13B, and 70B [20].

3.5 First Set of Tests

The following metrics were designed and justified based on the importance of the model: understanding natural language, providing logical responses, and remembering previous facts. The above considers that the objective is to generate a diagnosis and step-by-step guide for the implementation of Industry 4.0. Additionally, the models are expected to provide correct answers regarding the fundamentals of Industry 4.0 and demonstrate a high level of programming knowledge, as this is an aspect that involves all Industry 4.0 technologies.

Logical Reasoning (LR): The model is provided with a series of logic problems, such as paradoxes, inference questions, and logical sequences. The following questions are taken from [18]:

1. What relation do I have with my father's brother?
2. What weighs more, a kilogram of feathers or a kilogram of lead?
3. Juan's father has four children: Lucas, Sandra, Ana, and... who is the fourth?
4. Look at the following operation: 10 () 3 () 9 = 39. What are the correct operators?
5. If 5 machines make 5 items in 5 min, how long will it take 100 machines to make 100 items?

Expected answers are similar to the following:

1. He is your uncle
2. Both weigh the same, one kilogram
3. Juan is the four child
4. The operators to place are $\times$ and $+$ respectively, so the operation would be: $10 \times 3 + 9 = 39$
5. It will take 5 min since each machine makes one item every 5 min

Natural Language Understanding (NLU): The model is presented with complex text passages, then a series of questions are asked to assess its ability to understand and extract relevant information. Below is the first paragraph of the provided text, often used in academic texts or hobby sites:

"Hello! My name is Luis. When I was a child, my mother thought that young people should practice a sport and play an instrument. This encouraged me to play football and taught me teamwork, building confidence, and pursuing a common goal. It also motivated me to play the piano to improve coordination between my brain and body, cognitive ability, perseverance, and discipline..."

1. How many sports did Luis practice?
2. What were Luis's greatest fears?
3. What did music teach Luis?

Expected answers are similar to the following:

1. Four
2. Fear of dancing, swimming, and cooking
3. Luis developed a love for music, improved coordination between his brain and body, cognitive ability, and learned values like perseverance and discipline.

Code Understanding (CU): Code snippets in different programming languages are provided, and the model is asked to explain the function they perform. 1) What function does the code perform?, and the expected response: The provided code is an HTML page that displays a basic login form + "Explanation of the code.", (see Fig. 2).

```html
<!DOCTYPE html>
<html>
    <head>
     <title>Inicio de Sesión</title>
    </head>
    <body>
        <h2>Inicio de Sesión</h2>
        <form action="/login" method="post">
        <div class="container">
            <label for="uname"><b>Nombre de usuario</b></label>
            <input type="text" placeholder="Ingresa tu nombre de usuario"
              name="uname" required>
            <label for="psw"><b>Contraseña</b></label>
            <input type="password" placeholder="Ingresa tu contraseña"
                name="psw" required>
             <button type="submit">Iniciar Sesión</button>
        </div>
        <div class="container" style="background-color: #f1f1f1">
            <button type="button" class="cancelbtn">Cancelar</button>
            <span class="psw">¿Olvidaste tu <a href="#">contraseña?</a></span>
        </div>
        </form>
    </body>
</html>
```

Fig. 2. Code for login in HTML

Similarly, 2) a Python code snippet is provided, designed to find and print all prime numbers between 1 and 20. The code for resolving the factorial of a number is also presented, and it is asked: 3) What is the result of factorial (5)?

Code Generation (CG): The model is tasked with generating code that meets the required specifications. Write a Python function that takes a list as input and returns the list sorted in ascending order. The expected response should be similar to the showed (see Fig. 3).

```python
# Funcion
def ordenar_ascendente(lista):
    return sorted(lista)

# Ejemplo de uso:
mi_lista = [3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5]
lista_ordenada = ordenar_ascendente(mi_lista)
print("Lista original:", mi_lista)
print("Lista ordenada:", lista_ordenada)
```

Fig. 3. Function for ordering a list

Understanding of Technical Concepts (UTC): Questions are asked about the following concepts:

1. What is Industry 4.0?
2. What is the structure of a company?

Expected responses are similar to the following:

1. Industry 4.0, also known as the Fourth Industrial Revolution, refers to the integration of advanced digital technologies in manufacturing and production processes. It represents a significant evolution in how industrial processes are organized, managed, and executed...
2. The structure of a company refers to how it is internally organized to carry out its operations and achieve its objectives. While structures can vary depending on the type of organization, size, industry, and other factors, there are several common elements present in most companies. Here's a general description of the main components of a company structure...

Human-Machine Interaction (HMI): Tests the model's ability to maintain a conversation.

Human: Hello.

Possible AI response: Hi! How can I help you today?

Human: Today I went out and ate sushi at 2:15 PM at a restaurant where there's a riddle that says: "What color is Santiago's white horse?" Do you know the answer?

Possible AI response: Of course! The riddle "What color is Santiago's white horse?" has a well-known answer: the horse is white. It's a riddle that plays with redundancy to confuse a bit, but the answer is simply that the horse is white.

Human: Then I walked home and watched television.

Possible AI response: Sounds like a calm and relaxing day.

Human: Do you know what time I ate sushi?

Possible AI response: Yes, you mentioned you ate sushi at 2:15 PM.

3.6 Second Set of Tests

To carry out this second test, the Llama2 7B, Mistral 7B Instruct V0.2 and Gemma 7B models, all with 4-bit quantization were revised. The Spanish test implemented consists of five new questions focused on I4.0 technologies and their implementation.

1. What are the economic and operational challenges associated with the implementation of I4.0, and how can they be addressed to ensure the security, stability, and integrity of production processes?
2. How can workers adapt to the new demands of the labor market in a context of digitalization, where repetitive tasks are increasingly automated?
3. How can Denecke's (2008) methodology be used to conduct a precise sentiment analysis of tweets in Spanish, considering that they may contain multimedia content such as hashtags, mentions, and emojis?
4. From the following survey, provide a diagnosis of the I4.0 technologies implemented and identify which technologies are missing and would be beneficial for the company?
5. From the following survey, identify which I4.0 technologies are implemented in the surveyed company?, and make a list of the technologies that are missing for a good implementation. Additionally, give me a list of 10 steps to follow to gradually implement these and other technologies in the organization

Additionally, RAG was used, allowing the LLM to display responses with extremely recent data from three documents in Spanish about I4.0. The GPT4ALL tool was used, which allows running LLMs with or without RAG locally, something that LM Studio currently does not allow. The following documents were used in this process:

1. Current Situation of I4.0 [19]
2. Basic Aspects of I4.0 [4]
3. I4.0 from an Organizational Perspective [25]

Based on the project's data bank [7], a survey and artificial responses were generated using ChatGPT 3.5, indicating that the data corresponded to a company with wide gaps in I4.0 technology implementation.

1. Name: TextiTech Producers
2. Company size: Small
3. Annual sales volume: 10,000,000 to 50,000,000
4. Number of employees: 11 to 50
5. Knowledge of I4.0: Less than 25
6. Specific training: None
7. Barriers to the implementation of I4.0: High costs, Lack of human capital, insufficient infrastructure
8. Traceability: Traceability is not carried out
9. Budget for I4.0: Less than 1 total

10. I4.0 Technologies: No technological implementations specific logies
11. Cybersecurity: Conventional tools are used such as antivirus and firewall
12. Machine maintenance: Reactive maintenance
13. Sales channels and information collection: Networks social, physical catalogs
14. Main tool for data processing: Office

For questions 4 and 5 of the test, the model is presented the previous survey corresponds to the general characteristics of a company and I4.0 technologies. For the question 4 the object is for the LLM to respond correctly about a diagnosis of which technologies are present and which are not. Question 5 has important relevance since it requests a list of ten steps to gradually implement the I4.0 technologies. The list is based on the following works: [5] proposes a simple six-step road-map that includes real I4.0 implementations in "Small and medium-sized enterprise" (SME), and [2] provides small and medium-sized machine and installation manufacturers with a tool to develop their own business models and I4.0 implementations.

4 Results

This section evaluates the performance of the studied models. It was used the default configuration of the tools. Each answer to the questions posed to the models is scored according to the following criteria: **1** for a correct and appropriate answer to the question, **0** for a completely incorrect or senseless answer, and other values corresponding to incomplete answers. Here, the results are subject to the questions asked and represent an approximation of the models' behavior when interacting with them. Table 1 shows the disk size and memory usage of each model. The memory value was taken from the resource consumption log at the time of inference.

4.1 First Set of Tests

Table 2 shows the detailed results by model for the categories of questions asked in the first test, Gemma in the category of HMI had a result of zero because it argued aspects of information privacy. The three models achieved high performance in understanding of technical concepts (UTC), code understanding (CU)

Table 1. Resources used by the LLMs

Modelo	Quantization	Disk size (GB)	Memory size (GB)
Llama 2	Q2	2,83	3,6
	Q4	3,83	4,5
Mistral	Q2	3,08	3,0
	Q4	4,11	4,0
Gemma	Q2	3,48	4,2
	Q4	5,01	5,6

Table 2. Results of the first set of tests

Model	Q	LR	NLU	CU	CG	UTC	HMI	Points
English-Llama 2	Q2	3,5	2,7	2,0	1,0	2,0	1,0	12,2
	Q4	3,0	3,0	2,0	1,0	2,0	1,5	12,5
Spanish-Llama 2	Q2	0,0	2,7	2,0	1,0	2,0	0,0	7,7
	Q4	1,0	2,5	2,0	1,0	2,0	1,0	9,5
English-Mistral	Q2	2,0	2,0	2,0	1,0	2,0	2,0	11,0
	Q4	4,0	2,0	2,0	1,0	2,0	2,0	13,0
Spanish-Mistral	Q2	1,0	2,0	2,0	1,0	2,0	2,0	10,0
	Q4	3,0	2,0	2,0	1,0	2,0	2,0	12,0
English-Gemma	Q2	4,0	2,7	2	0,0	2,0	0,0	10,7
	Q4	2,0	2,7	2,0	1,0	2,0	0,0	9,7
Spanish-Gemma	Q2	2,0	1,7	1,8	1,0	1,8	0,0	8,3
	Q4	1,5	1,2	2,0	1,0	2,0	0,0	7,7
% Obtained		**45,0**	**75,6**	**99,2**	**91,7**	**99,2**	**47,9**	

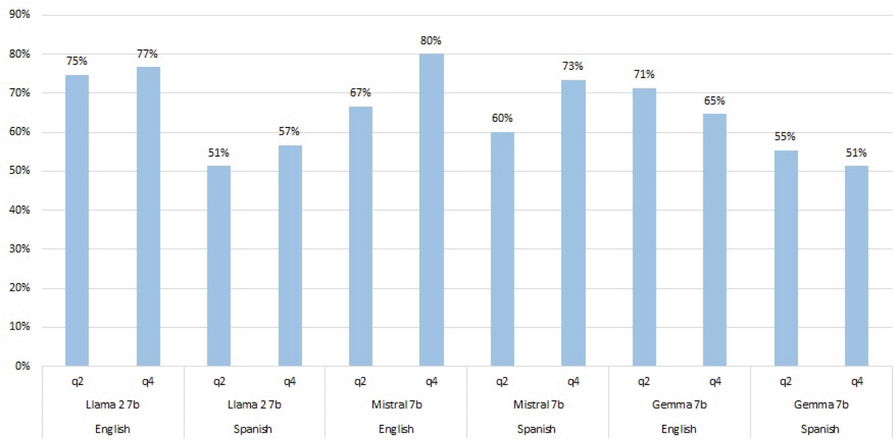


Fig. 4. Performance of the LLMs in the set first of tests

y code generation (CG). In natural language processing, Llama 2 achieved the best results for both English and Spanish.

The percentage of accuracy obtained by each model is shown, (see Fig. 4). Mistral with 7B and 4-bit quantization achieved the highest results in both cases, with 87% accuracy for the English case and 80% accuracy for the Spanish case. According to the results, it was not natural language understanding that contributed to Mistral's good performance as expected, but rather logical reasoning and the human-machine interface. For the Gemma model, accuracy decreased in both languages when moving from Q2 to Q4. LLama with q2 and q4 showed similar results for the English language.

4.2 Second Set of Tests

Table 3 shows the points obtained by each model with and without RAG from the answers to the questions in the second test. For the evaluation of the 5th question, a consolidated 10-step guide from documents [2,5] was used as a reference. It is not strict for the steps generated by the LLM to match exactly; they may vary depending on the organization where Industry 4.0 is intended to be implemented. In this question, the model that responded best was Mistral with RAG (75%), followed by Llama2 without RAG (74%), and the model with the worst response was Gemma without RAG, with only 18% accuracy. Mistral achieved 100% accuracy on the remaining questions.

Table 3. Results of the second set of tests

RAG	Model	1	2	3	4	5	Points	%
Without RAG	Llama 2	0,9	1	0,4	1	0,74	4,04	80,8
	Mistral	1	0,8	0,8	1	0,67	4,27	85,4
	Gemma	0,9	0,8	0,5	1	0,54	3,74	74,8
With RAG	Llama	1	0,5	0,5	1	0,53	3,53	70,6
	Mistral RAG	1	1	1	1	0,75	4,75	95
	Gemma-RAG	0,8	0,9	0,8	0,6	0,18	3,28	65,6

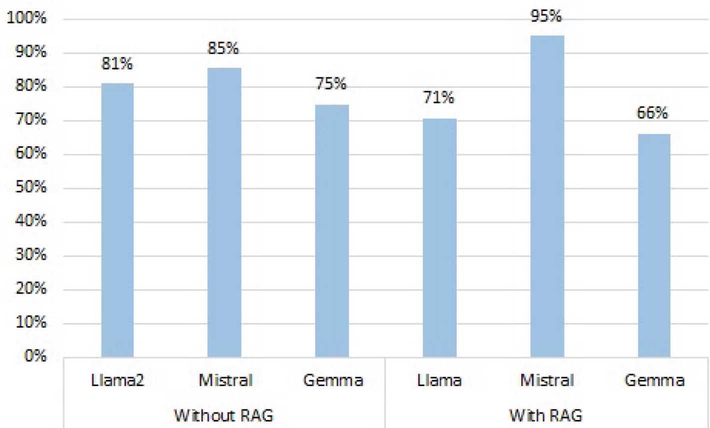


Fig. 5. Performance of the LLMs in the set second of tests

The percentage results achieved, where the maximum value than a model can reach is 5, equivalent to 100% (see Fig. 5). The Mistral model with RAG achieved the best performance with 95% accuracy and Gemma with RAG obtained the lowest accuracy, 66%. Gemma and Llama2 models did not improve their performance with the use of RAG. Mistral's high results in logical reasoning and the human-machine interface influence its strong performance when using RAG. Additionally, the integration of the Probabilistic Inference Layer (IPL) in Mistral leads to a marked improvement in the model's ability to understand the context and the intention behind the queries. This improvement is crucial to reducing tangential or irrelevant responses, thereby improving the model's overall effectiveness in natural language processing tasks [24].

5 Conclusions

A mid-range performance hardware setup is required for running the selected models without affecting performance. The response speed of each model depends on the level of quantization and the hardware on which it is running. For the first set of tests, each model took about 30 min to complete it. The model that achieved the best results in Spanish for all the tests was Mistral, compared to the Gemma and Llama2 models. In the first tests, for responses in English and Spanish, the best-performing model was Mistral, making it the most suitable choice for global applications. It was possible to demonstrate Mistral's enhanced ability to analyze the provided information and generate sufficient context to provide answers using RAG. In the second tests, with the use of RAG, this model significantly improved its results. Mistral fits the project's needs by accurately contextualizing the topic of I4.0 for which it was trained.

Acknowledgements. The authors thank the financial support provided through Call No. 890 of 2020 by Minciencias for the project 'Artificial Intelligence-Based Diagnosis of the Level of Industry 4.0 Adoption in Colombian Companies,' Contract No. 0665-2022.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. AI, M.: Open-weight models (2024). https://docs.mistral.ai/getting-started/open_weight_models/
2. Anderl, R., et al.: Guideline industrie 4.0-guiding principles for the implementation of industrie 4.0 in small and medium sized businesses. In: VDMA Forum Industrie, vol. 4, pp. 1–31 (2015)
3. Cloudflare: ¿qué es un modelo lingüístico (LLM) grande? (2024). https://www.cloudflare.com/es-es/learning/ai/what-is-large-language-model/
4. de las Tecnologías de la Información y la Comunicación, M.: Aspectos básicos de la industria 4.0 (2019). https://colombiatic.mintic.gov.co/679/articles-124767_recurso_1.pdf

5. Cotrino, A., Sebastián, M.A., González-Gaya, C.: Industry 4.0 roadmap: implementation for small and medium-sized enterprises. Appl. Sci. **10**(23), 8566 (2020)
6. Deng, H., Namoano, B., Zheng, B., Khan, S., Erkoyuncu, J.A.: From prediction to prescription: large language model agent for context-aware maintenance decision support. In: PHM Society European Conference, vol. 8, p. 10 (2024)
7. Fernández, J.M.L., Barrero, D.L.B., Rojas, L.A.R.: Industria 4.0: el reto para las pymes manufactureras de bogotá, colombia. Revista Mutis **12**(1) (2022)
8. Gallel, C.: Evaluación de los LLMs para la generación de código (2023)
9. Google: Modelos abiertos de gemma (2024). https://ai.google.dev/gemma?hl=es-419
10. He, L., et al.: Research status and application of artificial intelligence large models in the oil and gas industry. Pet. Explor. Dev. **51**(4), 1049–1065 (2024)
11. Hu, T., Zhou, X.H.: Unveiling LLM evaluation focused on metrics: challenges and solutions. arXiv preprint arXiv:2404.09135 (2024)
12. HuggingFace: Quantization (2024). https://huggingface.co/docs/optimum/en/con-cept_guides/quantization
13. IBM: What is fine-tuning? (2024). https://www.ibm.com/topics/fine-tuning
14. Javaid, M., Haleem, A., Singh, R.P.: A study on ChatGPT for industry 4.0: background, potentials, challenges, and eventualities. J. Econ. Technol. **1**, 127–143 (2023)
15. Kagglepro: Gemma vs llama vs mistral: a comparative analysis with a coding twist (2024). https://medium.com/@kagglepro.llc/
16. Lee, J., Su, H.: A unified industrial large knowledge model framework in smart manufacturing. arXiv preprint arXiv:2312.14428 (2023)
17. López-Riobóo-Botana, I., Gallent-Iglesias, D., Gonzalez-Vázquez, S.: Qua4i: question answering for the industry 4.0 domain. an application of intelligent virtual assistants (2023)
18. Martínez, C.: 50 preguntas de lógica con respuesta (2024). https://www.mundodeportivo.com/uncomo/ocio/articulo/50-preguntas-de-logica-con-respuesta-50852.html
19. de Mendoza, O.I.: Situación actual de la industria 4.0. https://bdigital.uncu.edu.ar/objetos_digitales/18983/oim-situacionactualindustriacuatrocero.pdf
20. Meta: Llama 2: open source, free for research and commercial use (2024). https://llama.meta.com/llama2/
21. Narula, S., et al.: Testing the accuracy of modern LLMs in answering general medical prompts. Int. J. Soc. Sci. Econ. Res. **8** (2023)
22. Ono, K., Morita, A.: Evaluating large language models: Chatgpt-4, mistral 8×7b, and google gemini benchmarked against mmlu. Authorea Preprints (2024)
23. Vidgen, B., et al.: Simplesafetytests: a test suite for identifying critical safety risks in large language models. arXiv preprint arXiv:2311.08370 (2023)
24. Wang, B., Wang, S., Ouyang, Q.: Probabilistic inference layer integration in mistral LLM for accurate information retrieval (2024)
25. Yandar Lobón, M.A., Moreno Ospina, J.M.: La industria 4.0: desde la perspectiva organizacional (2019). https://alinin.org/wp-content/uploads/2020/07/La-industria-40.pdf
26. Zeichick, A.: ¿qué es la generación aumentada de recuperación (rag)? (2023). https://www.oracle.com/es/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/

Optimal Integration of PV Sources and D-STATCOMs in Unbalanced Distribution Networks to Minimize Energy Losses: A Master-Slave Optimization Approach

Laura Sofia Avellaneda-Gomez[1(✉)], Brandon Cortés-Caicedo[1,2], and Oscar Danilo Montoya[1]

[1] Facultad de Ingeniería, Universidad Distrital Francisco Jose de Caldas, Bogotá 110231, Colombia
{lsavellanedag,bcortesc,odmontoyag}@udistrital.edu.co

[2] Facultad de Ingeniería, Institución Universitaria Pascual Bravo, Medellín 050036, Colombia
brandon.cortes@pascualbravo.edu.co

Abstract. This document evaluates a master-slave methodology developed to address the problem regarding the reduction of energy losses in electrical distribution systems by integrating PV generators and distribution static compensators. The master stage employs several optimization techniques, including the vortex search algorithm, particle swarm optimization, the Chu & Beasley genetic algorithm, and the generalized normal distribution optimizer, all aimed at determining the optimal location and size of each component. On the other hand, the slave stage uses the three-phase version of the successive approximations method to evaluate the objective function. To validate the efficacy of this approach, it was tested on a 15-node and a 35-node system. This study showcases an effective approach to minimize energy losses by simultaneously integrating two elements into the distribution grid. The findings indicate that the vortex search algorithm outperforms the other three optimization methods, reducing the energy losses by 58.65751% in the 15-node system and 71.9151% in the 35-node system. All numerical validations were carried out in the Julia software, version 1.10.2.

1 Introduction

Electrical distribution systems have grown rapidly due to population and industrial expansion [1]. This has led to the increased implementation of equipment that generates active and reactive power near the end-users [2–4]. Therefore, it is essential to explore alternatives that benefit grid operators. Notably, integrating photovoltaic (PV) generators and static compensators (D-STATCOMs) stands

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 41–53, 2025.
https://doi.org/10.1007/978-3-031-88854-0_4

out as a promising option [5], as these elements control active and reactive power flows, improve voltage profiles, and reduce network power losses [6].

Real distribution systems often exhibit an unbalanced operation due to end-user behavior and the geometric characteristics of distribution lines. This entails issues such as increased power losses, network segment overloading, service interruptions, and higher operating costs [7]. Additionally, implementing PV generators and D-STATCOMs in these systems implies significant planning and operational challenges [8]. These challenges have to do with the variability and intermittency of the solar resource in the case of PVs, the control of these devices, their compatibility with current grids designed for centralized generation, and the efficient management of active and reactive power [9], which can affect the reliability and stability of the system if done incorrectly.

Therefore, several approaches have been proposed in the literature to solve the problem regarding the optimal integration of PV generators and D-STATCOMs in power systems. Some of these are discussed below.

The authors of [10] discussed the main challenges and strategies associated with the integration and operation of distributed energy resources in three-phase systems. They highlighted the growing energy demand and the need for a transition towards more sustainable energy sources, which has led to increasing interest in optimizing the operation of electrical distribution grids.

The work by [11] presented a methodology for locating D-STATCOMs with the aim of minimizing power losses and improving voltage profiles and stability in a distribution system via the power losses minimization index, the current carrying capacity limit, and the voltage deviation index. These metrics were evaluated through simulations and statistical analysis, aiming to determine the overall performance enhancements achieved by the proposed solution.

The authors of [12] discussed the benefits of using PV generators, their modes of operation, their location in electrical systems, and the constraints to be considered for their siting and sizing. Among the observed benefits were the mitigation of environmental impacts, reduced power losses, and increased system capacity, among others.

As for the integration of D-STATCOMs, the literature reports a widespread use of analytical, metaheuristic, and hybrid algorithms to solve the PV generator siting and sizing problem while considering various objective functions. For instance, in [13], the simultaneous integration of reactive compensation and distributed generators was reported to increase power quality, reduce losses, increase system stability, provide operational flexibility, optimize costs, and promote environmental sustainability by incorporating renewable energy sources. The work by [14] integrated PV-STATCOM devices into the electrical grid to dynamically compensate for active and reactive power, improve voltage profiles, and reduce the annual energy losses.

Furthermore, in [15], the authors integrated PV generators by means of PSO, with the purpose of minimizing active power losses, improving voltage stability, and reducing the overall system imbalance. Their methodology considered seasonal fluctuations in load demand and the stochastic nature of PV generation, and it aimed to optimize the location of PV resources in distribution networks while simultaneously addressing multiple conflicting goals.

Three-phase networks have undergone several changes over time, an example of which is the integration of distributed generation, which makes the grid more complex in terms of planning and operation. In this vein, the work by [16] demonstrated how the use of metaheuristic algorithms allows solving the engineering problem posed by the adequate integration of elements into the network, considering that PV and D-STATCOM siting may be regarded as a variant of the optimal power flow problem. The authors of [17] discussed the benefits and challenges of integrating new elements into current distribution networks. They highlighted that proper power flow modeling and network analysis is essential to ensure a safe operation. Moreover, they compared four three-phase unbalanced power flow formulas to obtain the optimal power flow solution, with the purpose of proposing an alternative for adequately managing distribution grids and to stress the need to further explore this field.

Given the importance of integrating these devices into electrical distribution systems, this article proposes a strategy for locating and sizing PV generators and D-STATCOMs in unbalanced distribution networks while aiming to minimize energy losses. To this effect, a master-slave methodology based on metaheuristic optimization techniques is proposed. The master stage is responsible for defining the locations and sizes of the elements, for which four population-based techniques are implemented: the vortex search algorithm (VSA), particle swarm optimization (PSO), the generalized normal distribution optimizer (GNDO), and the Chu & Beasley genetic algorithm (CBGA). On the other hand, the slave stage evaluates the objective function and the constraints of the mathematical model for each of the individuals provided by the master stage. It uses a three-phase power flow tool based on the successive approximations method. The methodology was validated on test systems with 15 and 37 nodes, with the VSA yielding the best results.

The remainder of this document is organized as follows. Section 2 details the mathematical model representing the PV generator and D-STATCOM siting and sizing problem. Section 3 describes the implemented master-slave methodology. Section 4 outlines the test systems and provides additional considerations to validate the proposed methodology. Finally, Sect. 5 details the results obtained, and Sect. 6 presents the main conclusions of this work.

2 Mathematical Formulation

The integration problem concerning PV generators and D-STATCOMs in unbalanced distribution systems can be mathematically expressed as a mixed-integer nonlinear programming (MINLP) model, as it combines binary and continuous variables [5]. In this work, the minimization of energy losses was selected as the objective function, which is shown in Eq. (1). The MINLP model encompasses Eqs. (1) through (14).

$$\min\ E_{loss} = \text{Re}\left\{\sum_{p\in\mathcal{P}}\sum_{h\in\mathcal{H}}\sum_{k\in\mathcal{N}}\mathbb{V}_{k,h}^{p,*}\left(\sum_{j\in\mathcal{N}}\mathbb{Y}_{k,j}^{p}\mathbb{V}_{j,h}^{p}\Delta h\right)\right\}, \tag{1}$$

where $\mathbb{V}_{k,h}^{p}$ and $\mathbb{V}_{j,h}^{p}$ are the complex voltages of nodes k and j with phase p during period h; $\mathbb{Y}_{k,j}^{p}$ is the complex admittance connecting nodes k and j with phase p; Δh is the period of time during which the electrical quantities remain constant; and $\mathcal{N}$, $\mathcal{H}$, and $\mathcal{P}$ are the sets containing the nodes, periods, and phases of the system, respectively.

Set of Constraints:

Within the set of constraints, Eq. (2) ensures the complex power balance at all nodes, for all phases, and in all time periods. This equation considers the power injected by both the PV generators and the D-STATCOMs. Note that all PV generators, D-STATCOMs, and loads are assumed to have a Wye connection, which is the most common due to its advantages in terms of costs and protection [18].

$$\mathbb{S}_{k,h}^{gc,p} + \mathbb{S}_{k,h}^{pv,p} + \mathbb{S}_{k,h}^{ds,p} - \mathbb{S}_{k,h}^{d,p} = \mathbb{V}_{k,h}^{p}\left(\sum_{j\in\mathcal{N}}\mathbb{Y}_{k,j}^{p}\mathbb{V}_{j,h}^{p}\right)^{*}, \forall\begin{Bmatrix}k\in\mathcal{N},\\ h\in\mathcal{H},\\ p\in\mathcal{P}\end{Bmatrix} \tag{2}$$

where $\mathbb{S}_{k,h}^{gc,p}$, $\mathbb{S}_{k,h}^{pv,p}$, and $\mathbb{S}_{k,h}^{ds,p}$ denote, respectively, the complex power generated by the conventional generator, the PV generator, and the D-STATCOM at node k for phase p during time period h; and $\mathbb{S}_{k,h}^{d,p}$ is the complex power consumed at node k for phase p during time period h.

Equation (3) states that the complex power injected by the PV generators is evenly distributed among the phases of the system. The power injected by a PV generator depends on the resource availability and its nominal capacity.

$$\text{Re}\left\{\mathbb{S}_{k,h}^{pv,p}\right\} = G_{h}^{pv}\frac{\mathbb{S}_{k}^{pv}}{3}, \forall\begin{Bmatrix}k\in\mathcal{N},\\ h\in\mathcal{H},\\ p\in\mathcal{P}\end{Bmatrix} \tag{3}$$

where G_{h}^{pv} is the expected solar resource availability during time period h, and $\mathbb{S}_{k}^{pv}$ is the nominal complex power of the PV generator located at node k.

Equation (4) states that the PV generators inject only active power into the system.

$$\text{Im}\left\{\mathbb{S}_{k,h}^{pv,p}\right\} = 0, \forall\begin{Bmatrix}k\in\mathcal{N},\\ h\in\mathcal{H},\\ p\in\mathcal{P}\end{Bmatrix} \tag{4}$$

Equation (5) states that the D-STATCOMs inject only reactive power into the system.

$$\text{Re}\left\{\mathbb{S}_{k,h}^{ds,p}\right\} = 0, \forall\begin{Bmatrix}k\in\mathcal{N},\\ h\in\mathcal{H},\\ p\in\mathcal{P}\end{Bmatrix} \tag{5}$$

Equation (6) ensures that the complex power injected by the D-STATCOMs is evenly distributed among the phases of the system.

$$\operatorname{Im}\left\{\mathbb{S}_{k,h}^{ds,p}\right\} = \frac{\mathbb{S}_k^{ds}}{3}, \forall \begin{Bmatrix} k \in \mathcal{N}, \\ h \in \mathcal{H}, \\ p \in \mathcal{P} \end{Bmatrix} \tag{6}$$

where $\mathbb{S}_k^{ds}$ is the nominal complex power of the D-STATCOM located at node k.

Constraint (7) states that the conventional generators can only inject (*i.e.*, sell) power into the system, not absorb (*i.e.*, buy) it.

$$\operatorname{Re}\left\{\mathbb{S}_{k,h}^{gc,p}\right\} \geq 0, \forall \begin{Bmatrix} k \in \mathcal{N}, \\ h \in \mathcal{H}, \\ p \in \mathcal{P} \end{Bmatrix} \tag{7}$$

Constraint (8) specifies that the rated power of the PV generators to be installed must be within certain minimum and maximum limits.

$$x_k \mathbb{S}_k^{pv,\min} \leq \mathbb{S}_k^{pv} \leq x_k \mathbb{S}_k^{pv,\max}, \forall \left\{k \in \mathcal{N}\right\} \tag{8}$$

where $\mathbb{S}_k^{pv,\min}$ and $\mathbb{S}_k^{pv,\max}$ are the complex power limits of the PV generators at node k.

Constraint (9) specifies that the nominal power of the D-STATCOMs to be installed must fall within certain minimum and maximum limits.

$$y_k \mathbb{S}_k^{ds,\min} \leq \mathbb{S}_k^{ds} \leq y_k \mathbb{S}_k^{ds,\max}, \forall \left\{k \in \mathcal{N}\right\} \tag{9}$$

Here, $\mathbb{S}_k^{ds,\min}$ and $\mathbb{S}_k^{ds,\max}$ are the complex power limits of the D-STATCOMs at node k, and x_k and y_k are the binary variables associated with the location of the PV generators and D-STATCOMs, respectively.

Constraint (10) states that the voltage at the nodes in each phase must comply with regulations concerning voltage.

$$V_{\min}^p \leq \left|\mathbb{V}_{k,h}^p\right| \leq V_{\max}^p, \forall \begin{Bmatrix} k \in \mathcal{N}, \\ h \in \mathcal{H}, \\ p \in \mathcal{P} \end{Bmatrix} \tag{10}$$

where $V_{\min}^p$ and $V_{\max}^p$ are the upper and lower voltage bounds for the nodes in the system.

Constraints (11) and (12) define the maximum number of PV generators and D-STATCOMs that can be installed in the distribution system, respectively.

$$\sum_{k \in \mathcal{N}} x_k \leq N_{pv}^{ava}, \tag{11}$$

$$\sum_{k \in \mathcal{N}} y_k \leq N_{ds}^{ava}, \tag{12}$$

where N_{pv}^{ava} and N_{ds}^{ava} are constant parameters indicating the maximum number of PV generators and D-STATCOMs to be installed in the distribution system.

Finally, Constraints (13) and (14) specify that the variables x_k and y_k are binary.

$$x_k \in \{0, 1\}, \forall \left\{k \in \mathcal{N}\right\} \tag{13}$$

$$y_k \in \{0, 1\}, \forall \left\{k \in \mathcal{N}\right\} \tag{14}$$

Optimizing the placement and sizing of PV generators and D-STATCOMs in unbalanced distribution systems is challenging due to the problem's nonlinearities, the non-convexities in the power flow equations, and the combination of discrete and continuous variables [19,20]. This study employs methodologies based on a master-slave approach, which facilitates solving complex optimization problems, improves the search for optimal solutions, and reduces computation times [21].

3 Solution Methodology

In this work, due to the complexity of the problem, metaheuristic optimization techniques were selected instead of exact methods. Thus, a master-slave solution strategy was proposed. The master stage employed metaheuristic algorithms, *i.e.*, the VSA [22], the CBGA [23], the GNDO [24], and PSO [25], in order to determine the size and location of the PV generators and the D-STATCOMs. The slave stage used the three-phase version of the successive approximations method [26] to calculate the hourly load flow for each solution proposed by the master stage. This allowed evaluating the objective function.

Figure 1 summarizes the proposed master-slave methodology.

Master Stage: The master stage addresses the problem of integrating PV generators and D-STATCOMs into the distribution system. The main goal of the master stage is to identify the locations of the studied devices, along with their appropriate size, using a discrete-continuous encoding, as shown in [5].

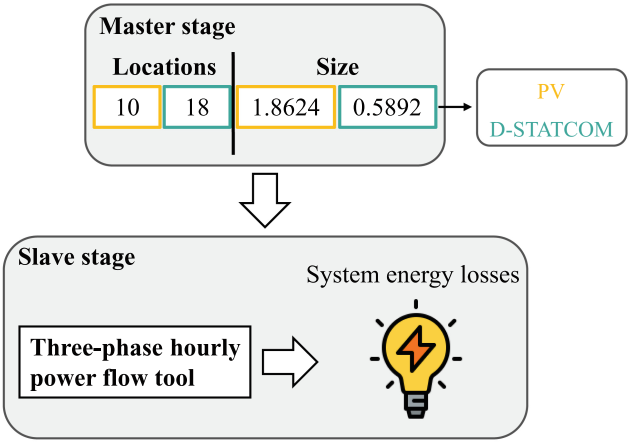


Fig. 1. Proposed methodology for the integration of PV generators and D-STATCOMs in unbalanced distribution networks

Slave Stage: The recursive power flow formula used in this document is shown below [26]:

$$\mathbb{V}_{d3\varphi,h}^{t+1} = -\mathbb{Y}_{dd3\varphi}^{-1} \left(\mathbb{Y}_{ds3\varphi} \mathbb{V}_{s3\varphi,h} - \left(\mathbb{I}_{d3\varphi,h}^{t} - \mathbb{I}_{pv3\varphi,h}^{t} - \mathbb{I}_{ds3\varphi,h}^{t} \right) \right) \tag{15}$$

where $\mathbb{V}_{d3\varphi,h}$ is the vector containing the three-phase voltage at the demand nodes in time period h; $\mathbb{Y}_{dd3\varphi}$ denotes the component of the three-phase admittance matrix that relates the demand nodes to each other; $\mathbb{Y}_{ds3\varphi}$ represents the component of the three-phase admittance matrix that relates the demand and generation nodes; $\mathbb{V}_{s3\varphi,h}$ signifies the vector containing the three-phase voltage at the generation nodes in time period h; $\mathbb{I}_{d3\varphi,h}$ is the vector containing the three-phase current at the demand nodes in time period h; $\mathbb{I}_{pv3\varphi,h}$ denotes the vector containing the three-phase current in time period h at the nodes where the PV generators are located; and $\mathbb{I}_{ds3\varphi,h}$ is the vector containing the three-phase current in time period h at the nodes where the D-STATCOMs are located.

4 Test Feeder Characterization

In order to validate the proposed methodology, it was applied to a 15-node test system (Fig. 2) and a 35-node feeder (Fig. 3). The former corresponds to a radial unbalanced distribution system with a nominal voltage of 13.2 kV at the terminals of the substation, in addition to a total active and reactive power consumption of 9,605 kW and 5,226 kvar for phase A, 6,480 kW and 4,940 kvar for phase B, and 11,977 kW and 8,778 kvar for phase C. The complete information regarding its electrical configuration, lines, and nodal power can be found in [27].

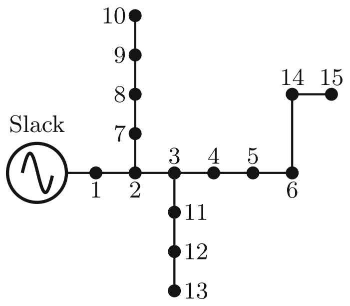


Fig. 2. Electrical configuration of the 15-node system

The 35-node test system is a radial unbalanced distribution system with a nominal voltage of 15 kV at the terminals of the substation. Its total active and reactive power consumption is 4,435 kW and 1,940 kvar for phase A, 2,868 kW and 2,059 kvar for phase B, and 2,925 kW and 1,370 kvar for phase C. The complete information of this test feeder can be found in [28].

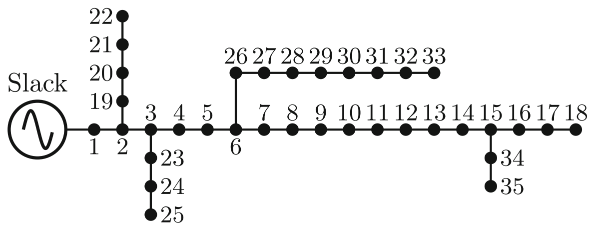


Fig. 3. Electrical configuration of the 35-node system

To evaluate the impact of installing PV generators and D-STATCOMs on the energy losses of these grids, generation and demand curves corresponding to the city of Medellín, Colombia, were employed (Fig. 4). These curves are further detailed in [29]. Additionally, it was assumed that the grid operator had access to three PV generators and three D-STATCOMs for installation. The nominal power of each device was determined through the proposed optimization methodology. However, to constrain the solution space, the maximum allowable power was set at 20,000 kVA per device. The size of the devices was related to the load level of both test systems, particularly for the 15-node system, which had a total consumption of 28,062 kW and 18,944 kVAr. Since the goal was to determine the optimal size and location of the studied devices while aiming to minimize energy losses, it became essential to impose the fewest possible constraints on their size. Also note that the investment costs of the PV generators and the D-STATCOMs did not influence the size selection.

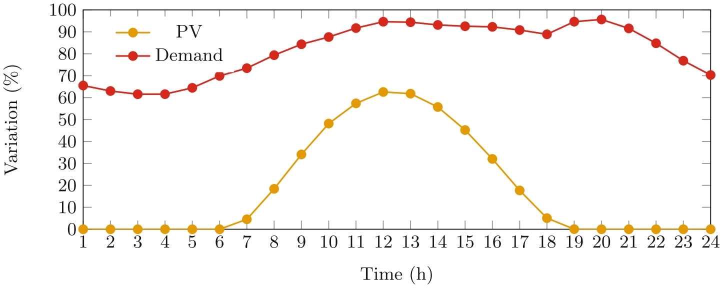


Fig. 4. Daily generation and demand curves for the city of Medellín, Colombia

5 Computational Validation and Results

Table 1 presents the simulation results for the 15- and 35-node test systems, detailing the methods used for siting and sizing the available PV generators and D-STATCOMs. Each method was evaluated based on the reported energy losses (kWh) and the percent reduction compared to the base case (*i.e.*, no optimization).

Table 1. Simulation results for the 15- and 35-node test systems

15-node test system				
Methods	PV (Node, MW)	DSTATCOM (Node, Mvar)	Energy losses (kWh)	Reduction (%)
Base case	–	–	1,815.0765	0.000
VSA	10(9.7146), 12(16.0892), 14(16.6351)	6(5.4914), 9(4.1861), 11(4.9059)	750.3981	58.6575
PSO	5(18.3077), 9(11.3791), 12(12.7502)	6(5.4773), 8(4.3816), 11(4.8773)	752.8707	58.5212
GNDO	9(14.3330), 12(11.4231), 14(16.6801)	6(6.5789), 8(3.3570), 11(3.5883)	762.6653	57.9816
CBGA	5(14.5500), 10(6.6144), 11(19.3562)	6(5.1703), 8(4.0617), 12(6.0580)	766.3478	57.7788
35-node test system				
Methods	PV (Node, MW)	DSTATCOM (Node, Mvar)	Energy losses (kWh)	Reduction (%)
Base case	–	–	13,725.6811	0.000
VSA	8(3.8674), 14(3.8199), 30(5.2805)	7(1.4637), 31(1.0153), 34(1.0548)	3,854.8482	71.9151
PSO	10(4.3127), 30(5.7355), 35(1.3298)	2(0.0000), 11(1.7020), 31(1.3042)	3,944.2490	71.2637
GNDO	11(6.3075), 18(0.4747), 30(5.2412)	9(1.5675), 17(0.3773), 31(1.2171)	3,957.2520	71.1690
CBGA	6(5.7520), 15(5.2264), 32(4.1791)	7(1.2052), 11(1.3471), 29(0.4493)	3,986.1767	70.9583

For the 15-node test system, the VSA was found to be the most effective method in reducing energy losses, achieving a value of 750.3981 kWh. This represents reductions of 1064.68 kWh (58.6575%) compared to the base case, 12.2672 kWh (0.6759%) compared to the GNDO, 15.9497 kWh (0.8787%) compared to the CBGA, and 2.4726 kWh (0.1363 %) compared to PSO. To achieve this solution, the VSA places the PV generators at nodes 10, 12, and 14, with a total installed capacity of 42.4389 MW. Additionally, the D-STATCOMs are placed at nodes 6, 9, and 11, with a capacity of 14.5934 MVAr.

For the 35-node test system, the VSA also stands out as the best methodology for reducing energy losses, with a value of 3,854.8482 kWh. This represents reductions of 9,870.833 kWh (71.9151%) compared to the base case, 102.4038 kWh (0.7461%) compared to the GNDO, 131.3285 kWh (0.9568 %) compared to the CBGA, and 89.4008 kWh (0.6514 %) compared to PSO. In this case, the VSA locates the PV generators at nodes 8, 14, and 30, with a total installed capacity of 12.9678 MW, and the D-STATCOMs are placed at nodes 7, 31, and 34, with a total capacity of 3.53381 MVAr.

Furthermore, the best methodology places the PV generators and the D-STATCOMs at nodes distant from the substation of both test systems, specifically in the outer branches. This arrangement allows the generation devices to be near the loads, which redistributes the power flow within the system and reduces the need to transport energy from the substation to the load nodes. Thus, a significant reduction in energy losses is achieved. The results indicate that the integration of PV generators and D-STATCOMs via metaheuristic techniques significantly reduces energy losses in three-phase distribution systems (by more than 58% with the VSA). The implementation of this approach in real systems can enhance their efficiency, decreasing the energy supplied by the substation and thereby reducing costs for the network operator. Moreover, the use of PV generators promotes the integration of renewable energies, contributing to sustainability and greenhouse gas emissions reduction efforts.

Table 2 compares the performance of the studied optimization methods after 100 consecutive evaluations in both test systems. For each method and test system, the following metrics are evaluated: best response, average response, worst response, standard deviation, and average processing time.

Table 2. Performance comparison for both test systems after 100 consecutive evaluations.

15-node test system					
Methods	Best (kWh)	Average (kWh)	Worst (kWh)	Std. (%)	Avg. time (s)
VSA	750.3981	770.6434	820.6519	2.4245	767.7245
PSO	752.8707	772.5653	826.4587	2.4723	723.2311
GNDO	762.6653	804.9596	847.3472	3.639	758.8070
CBGA	766.3478	790.6607	814.0341	1.7662	158.5052
35-node test system					
Methods	Best (kWh)	Average (kWh)	Worst (kWh)	Std. (%)	Avg. time (s)
VSA	3854.8482	3976.4527	4349.3894	3.7584	4405.4057
PSO	3944.2490	4191.7548	4807.5721	6.2376	3380.3818
GNDO	3957.2520	4134.3596	4480.3447	3.9215	4632.1364
CBGA	3986.1767	4023.1832	4364.9589	0.6553	758.8529

In terms of the average response, for the 15-node system, the VSA once again emerges as the most efficient methodology, achieving a value of 770.6434 kWh. This represents a reduction of 1.9219 kWh when compared to PSO, which ranks second in this regard. Similarly, regarding the worst solution, the VSA outperforms the other implemented methodologies. In the 35-node system, the VSA is also the most efficient methodology, achieving an average solution of 3976.4527 kWh and a worst response of 4349.3894 kWh, surpassing the other implemented methodologies in both aspects.

As for the standard deviation in the 15-node system, the CBGA reported the best score with 1.7662%, followed by the VSA with 2.4245%, the GNDO with 3.639%, and PSO with 2.4723%. The CBGA excels in this regard because its average is close to its extreme values (less than 30 kWh), indicating a tight clustering of the 100 solutions around the average. Among the methods used, the VSA performed best in relation to the average, best, and worst responses, with all metrics surpassing CBGA's best performance. The VSA's standard deviation ensures solutions close to its average, outperforming the other methods. In the 35-node system, CBGA ranked first with 0.6553%, followed by the VSA with 3.7584%, the GNDO with 3.9215%, and PSO with 6.2376%.

In terms of processing times, for the 15-node system, the CBGA yields significant improvements over the other methodologies. This algorithm averaged 158.5052 s to converge to a solution, notably faster than the GNDO, PSO, and the VSA, which averaged 758.807 s, 723.2311 s, and 767.7245 s, respectively. As for the 35-node system, the CBGA shows a marked improvement compared to the other methodologies; on average, it took 758.8529 s to converge, which is also significantly faster than the GNDO, PSO, and the VSA, (4632.1364 s, 3380.3818 s, and 4405.4057 s, respectively).

The results in Table 2 show the low variability of the VSA and the CBGA with regard to energy losses, indicating a consistent and reliable performance, which is crucial for the planning and operation of electrical grids.

6 Conclusion

This paper proposes a master-slave methodology for siting and sizing PV generators and D-STATCOMs in unbalanced distribution networks while aiming to minimize energy losses. Said proposal was validated using different optimization techniques (GNDO, CBGA, PSO, and VSA) in 15- and 35-node test systems. The VSA outperformed the CBGA, the GNDO, and PSO in terms of energy losses reduction, with a value of 58.6575% *versus* the 57.7788%, 57.9816%, and 58.5212% reported by the other methods, respectively. The VSA also demonstrated excellent repeatability conditions, allowing for solutions equal or very close to the mean value. This work showed that, in two test systems, the VSA is the most successful in identifying the best solutions regarding energy losses reduction.

As future work, the following studies could be conducted: (i) proposing a formulation that considers technical, economic, and environmental objective functions; (ii) applying new combinatorial optimization algorithms to validate or improve the numerical results presented in this document; and (iii) extending the proposed formulation to include battery energy storage systems.

References

1. Bansal, M., Garg, P., Prashar, A.: Recent advancement in electricity transmission and distribution technologies. Turkish Online J. Qual. Inquiry **12**(7) (2021)
2. Muhtadi, A., Pandit, D., Nguyen, N., Mitra, J.: Distributed energy resources based microgrid: review of architecture, control, and reliability. IEEE Trans. Ind. Appl. **57**(3), 2223–2235 (2021)
3. Satish, R., Vaisakh, K., Abdelaziz, A.Y.: Solving issues of grid integration of solar and wind energy models by using a novel power flow algorithm. In: Wind and Solar Energy Applications, CRC Press, 2023, pp. 251–266 (2023)
4. Baviskar, A., Hansen, A.D., Das, K.: Reactive power support from converter connected renewables in active distribution network. In: 2022 IEEE 13th International Symposium on Power Electronics for Distributed Generation Systems (PEDG), pp. 1–6. IEEE (2022)
5. Rincón-Miranda, A., Gantiva-Mora, G.V., Montoya, O.D.: Simultaneous integration of D-STATCOMS and PV sources in distribution networks to reduce annual investment and operating costs. Computation **11**(7), 145 (2023)
6. Deng, G., Wang, D., Gao, W.: Active and reactive power coordination optimization of the active distribution network. J. Phys. Conf. Ser. **2450**, 012023 (2023). IOP Publishing
7. Alvarado-Barrios, L., Alvarez-Arroyo, C., Escano, J.M., Gonzalez-Longatt, F.M., Martinez-Ramos, J.L.: Two-level optimisation and control strategy for unbalanced active distribution systems management. IEEE Access **8**, 197992–198009 (2020)
8. Fardinfar, F., Pour, M.J.K.: Optimal placement of D-STATCOM and PV solar in distribution system using probabilistic load models. In: 2023 10th Iranian Conference on Renewable Energy & Distributed Generation (ICREDG), pp. 1–5. IEEE (2023)
9. Deepthi, R., Saritha, M., Sidram, M.: Design and performance analysis of three phase solar PV grid integrated DSTATCOM, in: 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 213–218. IEEE (2019)
10. Guzmán-Henao, J.A., Bolaños, R.I., Montoya, O.D., Grisales-Noreña, L.F., Chamorro, H.R.: On integrating and operating distributed energy resources in distribution networks: a review of current solution methods, challenges, and opportunities, IEEE Access (2024)
11. Kandasamy, M., et al.: Strategic incorporation of DSTATCOM and distributed generations in balanced and unbalanced radial power distribution networks considering time varying loads. Energy Rep. **9**, 4345–4359 (2023)
12. Bawazir, R.O., Cetin, N.S.: Comprehensive overview of optimizing PV-dg allocation in power system and solar energy resource potential assessments. Energy Rep. **6**, 173–208 (2020)
13. Dash, S.K., Mishra, S., Abdelaziz, A.Y., Hong, J., Geem, Z.W.: Optimal planning of multitype DGS and D-STATCOMS in power distribution network using an efficient parameter free metaheuristic algorithm. Energies **15**(9), 3433 (2022)
14. Garrido-Arévalo, V.M., Gil-González, W., Montoya, O.D., Chamorro, H.R., Mírez, J.: Efficient allocation and sizing the PV-STATCOMS in electrical distribution grids using mixed-integer convex approximation. Energies **16**(20), 7147 (2023)

15. Maji, S., Kayal, P.: A simplified multi-objective planning approach for allocation of distributed PV generators in unbalanced power distribution systems. Renew. Energy Focus **48**, 100541 (2024)
16. Bai, W., Zhang, W., Allmendinger, R., Enyekwe, I., Lee, K.Y.: A comparative study of optimal PV allocation in a distribution network using evolutionary algorithms. Energies **17**(2), 511 (2024)
17. Mahmoodi, M., RA, S.M.N., Blackhall, L., Scott, P.: A comparison on power flow models for optimal power flow studies in integrated medium-low voltage unbalanced distribution systems. Sustain. Energy Grids Netw. **38** 101339 (2024)
18. Sepúlveda, S., Ruíz, A.G., Mora-Flórez, J.J.: Optimal power flow for three-phase unbalanced active distribution networks with delta connections. In: 2022 IEEE ANDESCON, pp. 1–6. IEEE (2022)
19. Aldik, A., Venkatesh, B.: Reactive power planning using convex line-wise power balance equations for radial distribution systems. IET Gener. Transm. Distrib. **14**(12), 2399–2406 (2020)
20. Liu, B., Li, J., Ma, H., Liu, Y.: Generalized benders decomposition based dynamic optimal power flow considering discrete and continuous decision variables. IEEE Access **8**, 194260–194268 (2020)
21. Devikanniga, D., Vetrivel, K., Badrinath, N.: Review of meta-heuristic optimization based artificial neural networks and its applications. J. Phys. Conf. Ser. **1362**, 012074 (2019)
22. Doğan, B., Ölmez, T.: A new metaheuristic for numerical function optimization: vortex search algorithm. Inf. Sci. **293**, 125–145 (2015)
23. Chu, P.C., Beasley, J.E.: A genetic algorithm for the generalised assignment problem. Comput. Oper. Res. **24**(1), 17–23 (1997)
24. Zhang, Y., Jin, Z., Mirjalili, S.: Generalized normal distribution optimization and its applications in parameter extraction of photovoltaic models. Energy Convers. Manage. **224**, 113301 (2020)
25. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN 1995-International Conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE (1995)
26. Montoya, O.D., Garces, A., Gil-Gonzalez, W.: Three-phase power flow tool for electric distribution grids: a Julia implementation for electrical engineering students, Ingeniería **28**(3) (2023)
27. Sudhakar, T., Rajan, M.R., Srinivas, K., Prabu, R.R., Narmadha, T., Krishnan, M.M.: Modeling and simulation of distribution network with the integration of distribution generator using MATLAB. Indian J. Sci. Technol. **9**(12), 1–7 (2016)
28. Montoya-Giraldo, Ó.D., Ramírez-Vanegas, C.A., González-Granada, J.R.: Power losses reduction in three-phase unbalanced distribution networks using a convex optimization model in the complex domain. TecnoLógicas **27**(59), e2903–e2903 (2024)
29. Cortés-Caicedo, B., Grisales-Noreña, L.F., Montoya, O.D., Rodriguez-Cabal, M.A., Rosero, J.A.: Energy management system for the optimal operation of PV generators in distribution systems using the antlion optimizer: a Colombian urban and rural case study. Sustainability **14**(23), 16083 (2022)

Duolingo Evolution: From Automation to Artificial Intelligence

Janice Vega[1(✉)], Monica Rodriguez[1], Erick Check[2], Halpin Moran[2], and Luis Loo[3]

[1] Dept. Languages and Linguistics, Southern Illinois University, Carbondale, USA
{mirna.vegalacayo,monicaisabel.rodriguezherman}@siu.edu
[2] Dept. Maestria U.T.H, Universidad Tecnologica de Honduras, San Pedro Sula, Honduras
[3] Dept. Posgrados U.T.H, Universidad Tecnologica de Honduras, San Pedro Sula, Honduras
luis.loo@uth.hn

Abstract. This paper investigates the transformative evolution of the Duolingo language learning application, focusing on its shift from automation to Artificial Intelligence (AI) integration. Through a comprehensive secondary data analysis, we explore how Duolingo has adapted its platform to incorporate AI-driven tools, such as Automatic Speech Recognition (ASR) and machine learning models, to enhance user experiences with personalized, real-time feedback and interactive learning. The paper traces the historical development of Computer Assisted Language Learning (CALL) and its alignment with technological advancements, highlighting Duolingo's migration from Phyton to Scala, which significantly improves scalability, stability, and performance. Additionally, this study delves into the challenges faced during this migration, such as technical debt and compatibility issues, while emphasizing the strategic decisions that shaped Duolingo's current architecture. The findings illustrate how AI functionalities, like spaced repetition algorithms, and GPT'powered features, have redefined the language learning landscape by offering adaptive, efficient, and engaging educational experiences. Ultimately, Duolingo's continuous innovation serves as a model for the future of technology-enhanced language learning in an increasingly digital age.

Keywords: Duolingo · Automation · Artificial Intelligence (AI) · Computer Assisted Language Learning (CALL) · evolution · pedagogical shifts · Python · Scala · language learning apps

1 Introduction

This paper aims to provide evidence of the significant technological evolution of the Duolingo app has undergone, transitioning from an automated system to one deeply integrated with Artificial Intelligence (AI). As a case study, Duolingo offers a unique perspective on the intersection of CALL (Computer Assisted Language Learning), pedagogical shifts, and rapid technological advancements. The evolution of CALL, influenced by the behaviosris approach of Skinner in the mid-20th century, has largely mirrored technological progress, beginning with simple automation and eventually incorporating complex AI systems such as Atuomatic Speech Recognition (ASR), natural language

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 54–71, 2025.
https://doi.org/10.1007/978-3-031-88854-0_5

processing (NLP), and machine learning algorithms to improve the efficacy of language learning experiences.

Duolingo's transition exemplifies this evolution, showing how the app has gradually integrated AI capabilities to improve its functionality and effectiveness in facilitating language learning [1] Specifically, Duolingo's migration from Python to Scala is explored as a critical turning point, allowing the app to streamline its architecture, optimize computational efficiency, and better accommodate AI-driven functionalities like personalized feedback, real-time corrections and adaptive learning paths that cater to individual learner needs [2].

However, such integration presents inherent challenges, including the complexities of migrating between programming languages, addressing technical debt, and the critical need to safeguard user data from security breaches without compromising app functionality [2].

2 Research Questions

The advancement of Computer Assisted Language Learning (CALL) has greatly impacted the methods and tools utilized by language learning platforms. Considering that CALL has continuously adjusted to new developments in CALL and achieved learning, this change demonstrates a fluid evolution within the technology used for language education.

Duolingo, as a prominent example, has experienced this shift by gradually integrating AI capabilities into its platform. However, how this evolution impacts the transition of Duolingo from automation to AI remains a crucial area of inquiry. For this reason, this paper investigates the underlying factors that have driven Duolingo's transition from automated procedures to the integration of AI functionalities and how they convey this process. In addition to the shift happening between automation and Artificial Intelligence, our inquiry delves into Duolingo's automation features this application keeps hold of during this transformation process. Additionally, we endeavor to uncover the impact of the pedagogical landscape changes on the evolution of Duolingo innovations. To explore these questions, a secondary data analysis of Duolingo's developmental trajectory takes place, examining historical milestones and strategic decisions that have shaped its evolution.

By investigating the relationship between CALL and technological advancements, and pedagogical approaches, we aim to uncover the underlying motivations urging Duolingo's transition toward AI integration. Additionally, the paper explores how Duolingo uses automation tools, such as Automatic Speech Recognition (ASR), to enhance user experiences, AI functionalities, and facilitate language learning. For this reason, this paper seeks to contribute to a deeper understanding of the complex dynamics driving the evolution of language learning platforms in the digital age.

3 Historical Background for CALL

According to Chapelle and Sauro (2017) [3] first-generation computers, primarily operated through paper punch cards and began to appear on college campuses during the late 1950s. However, it was not until the late 1960s and early 1970s that computers had evolved sufficiently to accommodate multiple terminals, enabling interaction with users through keyboards. This development marked the emergence of Computer Assisted Language Learning (CALL).

Kumar and Sreehari (2009) [4] defined Computer Assisted Language Learning, also known as CALL, as "an interactive method of instruction that helps learners achieve their goals of learning, at their own pace and ability. In this method, computer technology is used in teaching and learning procedures at all stages such as presentation, practice, and feedback." This definition aligns with this paper's context as it acknowledges the integration of mobile applications such as Duolingo, allowing students to progress at their own speed while enhancing language skills.

It also underscores the evolution towards utilizing convenient technological tools, transitioning from earlier automatic tools embedded within learning software to today's incorporation of Artificial Intelligence. With the emergence of each new technology, we undergo a continual cycle of assessing and improving our reasoning and approaches to its implementation. At the core of this cycle, we can find the fundamental concerns that Clarke highlighted:

Quality of materials, teacher engagement and training, suitability of technology to specific instructional goals, and adherence to pedagogical principles [3] However, the language learning settings before this shift mainly focused on the teaching of grammatical rules which were typically learned through translation between the native language and the target language. As technology progressed and computers became more prevalent, they were initially perceived as rudimentary tutors of the target language with limited capabilities. Consequently, the majority of CALL technologies involved exercises that primarily focused on grammatical activities such as vocabulary flashcards for memorization, text-based grammar drill-and-practice devoid of context, and translations [3].

Nonetheless, while both pedagogical and technological advancements of the time were under deep analysis, criticism arose towards the predominant use of CALL tools for grammatical activities, as it was seen to lack contextual understanding [5]. This led to new and revolutionary pedagogical trends, with one particularly significant trend emerging in the late 1980s: the focus on establishing an authentic environment for language learning. This shift marked a departure from the method of Grammar Translation towards the adoption of Communicative Language Teaching (C.L.T.). This methodology can be defined as a framework for language teaching that prioritizes the development of communicative competence [6].

Communicative competence encompasses the ability to effectively utilize the language across various purposes, contexts, and interlocutors, among other factors. This evolution of pedagogical methods was inevitably integrated into CALL, having the internet as an important role in it. This development saw notable advancements such as the use of software for information retrieval, the introduction of the World Wide Web (referred to as Web 1.0), and the ability to incorporate multimedia resources in educational environments [5]. Consequently, the internet opened up broader avenues for accessing authentic language learning materials, environments, and interactions with people.

This period marked a transition for CALL towards an integrative phase, where tasks aimed at immersing learners in an authentic context and integrating various language skills for their development [5] Currently, with the evolution of language learning methods and approaches, the same as technology emerging, it is widely acknowledged that the use of technology in modern education is essential to keep pace with society's advances. The globalized world continuously develops more global technology, to connect and develop more knowledge, science, and vast information. This integration of technology enhances teaching and learning but also ensures that we remain relevant and competitive in today's fast-evolving landscape [7].

4 Duolingo Language Learning App

Duolingo is a learning phone app that is available on Apple Store and Play Store where students can download it on mobile devices the same as on a desktop computer and then start learning a language; it may be used with or without an internet connection [8]. Duolingo was launched as a private beta on November 27, 2011, with six languages by their creators Luis von Ahn and Severin Hacker. Later on, they opened Duolingo to the public on June 19, 2012. According to Adams S. (2019) [9] Forbes magazine claimed that the app's three-minute lessons are designed with a simple interface that can be used by the over 300,000 users that have downloaded the app by 2019. Moreover, it was first developed in the programming language Python, integrating technological components with high-quality content and algorithms that underpin its success.

Nonetheless, part of the evolution of the app as with any other one is the step back in its technological development process. In 2014 they attempted to introduce a virtual chatbot tutor named Duo. Its function was to simulate a conversation with a native speaker, but the chatbot was found ineffective and lacking in nuance [10].

Additionally, according to Meyer (2021) [11] many contemporary learning materials designs frequently blend visual elements, such as pictures or graphics, along with text information, whether presented in a spoken or written way. These integrated combinations are commonly known as multimedia learning materials, which are part of what makes Duolingo a successful learning app. For instance, not only his official mascot Duo the Owl, bond with the users of the app and keep them engaged with their learning language process, but also " Duolingo grabs users with gamification tricks like points, treasure chests and "streaks" for continuous use"[9].

Regarding users' experiences with these recent enhancements to Duolingo, a survey [1] conducted among participants who have utilized Duolingo for less than one year revealed that the platform is highly comprehensible and its overall quality is rated as notably high.

5 Discussion

A. Factors that have Driven Duolingo's Shift to AI

In the last decade, research on Computer Assisted Language Learning (CALL) has increasingly focused on understanding the cognitive process involved in language acquisition, particularly how learners process stimuli through their working memory, known as cognitive load [12].

Cognitive Load Teaching (CLT), which aims to optimize learning by presenting information in a way that aligns with human cognitive capabilities, has been foundational in shaping Duolingo's approach to language learning. By recognizing the limitations of working memory and its interaction with long-term memory, Duolingo's designs reflect principles of CLT, particularly in how structures learning activities to maximize cognitive overload [13].

Moreover, one of the main factors driving Duolingo's shift to Artificial Intelligence (AI) is the platform's need to enhance cognitive load management in real time. Ai provides adaptive learning experiences that tailor the content and difficulty based on the learner's performance, allowing users to focus on learning without being overwhelmed by too much information at once. For example, Fuolingo's Half-Life Regression model, grounded in cognitive load theory, helps reduce the burden on working memory by predicting when users are likely to forget material and scheduling reviews accordingly.

This algorithm ensures that learners review content at the optimal time, increasing retention and reducing cognitive fatigue. Furthermore, AI's ability to assess performance instantaneously and adjust learning paths accordingly ensures that learners are not exposed to unnecessarily complex tasks. As learners progress, AI-driven systems adapt the content's complexity, providing a scaffolded learning environment that promotes gradual mastery of language skills without exceeding cognitive limits.

Another factor influencing the shift to AI has been the need for scalability and personalization. As Duolingo's user base has grown, it has become essential to deliver highly individualized learning experiences at scale. AI has made this possible by automating much of the content creation process while still ensuring personalized feedback.

For instance, AI-powered tools like Birdbrain optimize the difficulty and sequence of lessons, making the learning process more efficient while ensuring that each learner's cognitive load is effectively managed. Duolingo shift to AI is deeply influenced by both technological advancements and cognitive learning principles. The use of AI not only helps menace cognitive load but also allows Duolingo to deliver personalized, scalable, and adaptive learning experiences that are aligned with the latest research in CALLA and CLT [12, 13].

B. Automation of Duolingo App

The automatization process of language learning started with the integration of ASR (Automatic Speech Recognition) technology, these systems are also known as Computer-Assisted Pronunciation Training (CAPT), which may have provided limited interaction: the computer understands the student's speech and reacts accordingly, thus making the learning process more realistic and engaging, and provides feedback on the quality of the student's speech [14].

This is often confused with Artificial intelligence, however, the main difference here is that automatic technology has been programmed to respond to specific stimuli, if there is anything outside this specific stimuli, no accurate response can be produced by the technology itself. The optimal ASR-based CAPT system in 2003, was conceptualized as a series of five phases: Speech recognition, scoring, error detection, error diagnosis, and feedback presentation [14].

The latter three elements are now commonly integrated into popular artificial intelligence tools such as Duolingo, Eliza, Fluent U, and Gliglish among others, which offer immediate feedback based on the learners' mistakes. The following outlines the five key stages of the process of automated language assessment:

(a) Automatic Speech Recognition (ASR) stands as the initial phase, where human speech undergoes processing based on predefined rules and patterns to ascertain its correctness. Advanced ASR applications, like those integrating artificial Intelligence learning, incorporate grammar, syntax, and structural components within audio and voice signals to redefine speech processing. Consequently, AI-powered ASR systems learn and improve alongside learners' interactions [15].
(b) Automated scoring (AS) on its end, involves translating students' tasks into assessments of quality. This process is driven by motivations such as cost reduction, scalability, immediate feedback provision, and ensuring assessment consistency and accuracy [16].
(c) Error detection plays a pivotal role in enhancing learners' accuracy by identifying potential transcription errors within the output of ASR systems. Leveraging word-level confidence scores from the ASR system, error direction performance can be further augmented, for instance, through mechanisms like the ASR Confidence Embedding (ACE) layer in the Error Detection Model (EDM) encoder [17].
(d) The error diagnosis phase, aligning with scoring and error detection, utilizes algorithmic approximations of human rating to match programmed rubrics [16].
(e) Feedback presentation encompasses task scores along with individual linguistic features, which are generally perceived as beneficial by teachers [16].

C. Duolingo's Evolution from Automation to Artificial Intelligence

Yousefpoori-Naeim [18] emphasized the innovative and popular nature of the Duolingo language learning app. The app's continuous adaptation to technological advancements and educational innovation plays a pivotal role in its increasing popularity. Firstly, this app transitioned its lesson engine from Python to Scala, leading to enhancement in performance, stability, and uptime. This shift effectively addressed technical debt and scalability concerns, simplifying the architecture and capitalizing on Scala's functional programming capabilities.

The chosen technology stack included Finatra, Guice, and Mockito, facilitating extensive testing and seamless deployment on AWS Elastic Beanstalk. Although encountered challenges included documentation gaps and compatibility issues with Java libraries, the rewrite notably sashed latency (from 750ms to 14ms) and error rates while preserving performance levels [2].

Despite the migration process to Scala, it is still useful to fully define Phyton as a programming language. Phyton is a widely-used, high-level programming language that is both interpreted and object-oriented, offering dynamic capabilities. Its appeal lies in its combination of dynamic typing and binding alongside robust built-in data structures, making it particularly suitable for fast application development and scripting tasks. This versatility allows Python to easily integrate different components of larger systems.

The language's syntax is designed to be simple and readable, which significantly reduces the complexity of writing and maintaining code. Furthermore, Python promotes molarity and code reuse through its support for modules and packages. Its interpreter, along with an extensive standard library, is freely available in both source and binary formats for all major platforms. Python's accessibility and flexibility have made it an indispensable tool across a variety of fields, from academia to industry [19].

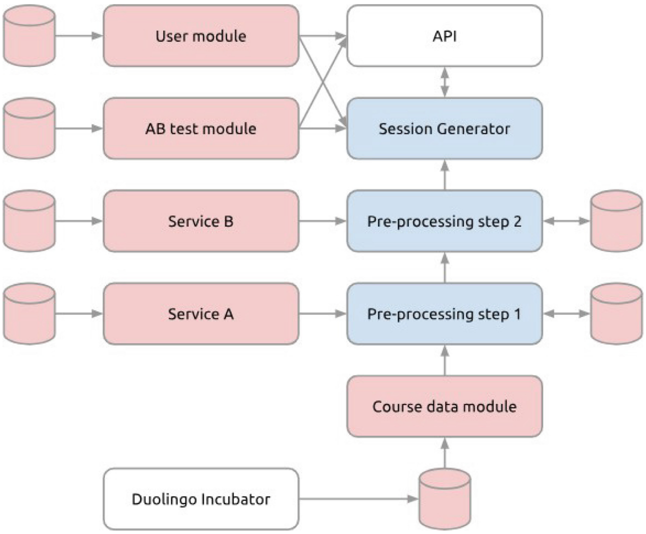


Fig. 1. Phyton's Architectural Design [2].

The architecture shown in the flow diagram above (Fig. 1) describes a system consisting of four primary server connections: The user model, AB test module, service B, and service A. Each server connection is linked to at least one other function, with some requiring decision points, such as in user connection where the next step could either involve an API call or session generation and vice-versa with the AB test module. The main challenge this architecture type represents is the rigid dependencies on the server connection and the primary app connections, these additional network calls contribute to the overall latency of the system. For instance: "Session Generator would fail if any of these (or their respective network connections) failed" [2].

The transition of this architecture to a different one was prompted by several factors, including sociological changes, increased user access to technologies, and the accumulation of technical debt with the application. As per Horie, A. K (2020) [2].

Technical debts refer to the practice of addressing shortcuts discovered through app usage that ultimately require rectification to align with technological development and user preferences. Technical decisions were made at first to establish a system architecture capable of expediting development while supporting application features. The same as Phyton, Scala needs to be defined to deeply understand the features that led to migration.

Scala is a modern, multiparadigm language designed to express common programming patterns with clarity, brevity, and type safety. It uniquely blends object-oriented as functional programming features, providing a robust and versatile tool for developers. As an object-oriented language, Scala treats every value as an object, with types and behaviors defined by classes and traits. It supports subclassing and a flexible mixin-based composition mechanism, offering a cleaner alternative to multiple inheritance.

On the functional side, Scala treats functions and supports higher-order functions, function nesting, and currying. Its case classes and pattern matching offer functionality akin to algebraic types, a common feature in many functional languages. Additionally, Scala's pattern-matching capabilities extend to processing complex data structures like XML, making it highly suitable for applications such as web services. These features, combined with the ability to group standalone functions into singleton objects, position Scala as a powerful language for both object-oriented and functional programming tasks [20].

The new architectural design presented in the flow diagram above (Fig. 2) encompasses only 3 main connections: the server user module, AB test module, and services. Both, the user module and AB module services connect to an API, which then progresses to the session generator, written in Scala, before reaching the server. It is noteworthy to mention, that the course data architecture requires more processing steps, but as it is shared across users in the same course, it is processed offline and utilizes files in AWS S23, Amazon's cloud storage service [2].

Redesigning the engine system was found as the most effective way to minimize shared resources. This resulted in a simpler design, even when it required a longer codebase adaptation period due to the lack of full compatibility with Python programming language. Consequently, there was a decision to migrate to the Scala programming language, starting by rewriting Session Generator in Scala [2].

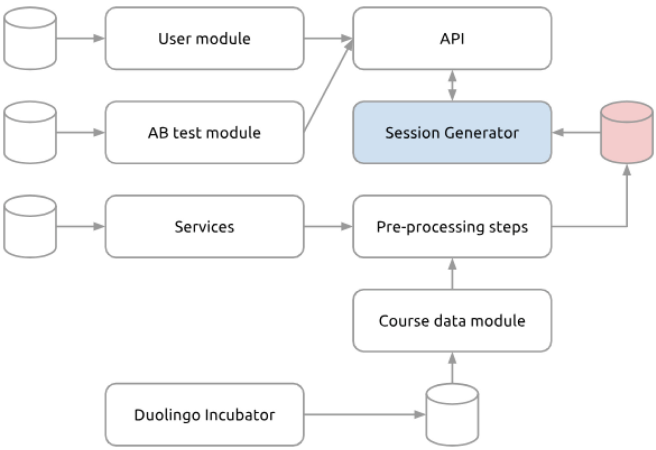


Fig. 2. Scala's Architectural Design [2].

As shown (Fig. 3), Duolingo has evolved from using Phyton to Scala and later advanced to utilize generative AI with GPT-3 and GPT-4. This paper only delves into the

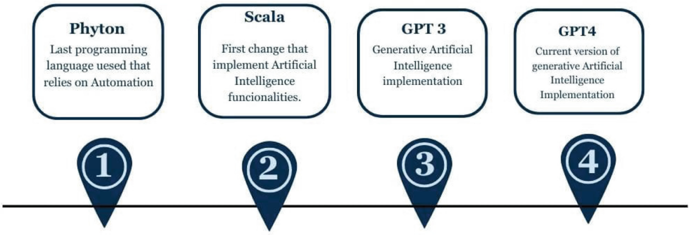


Fig. 3. Duolingo Evolution From Automation to Artificial Intelligence [21].

transition process from mere automation to the implementation of artificial intelligence, focusing specifically on the transition from Python to Scala.

D. Improvement of AI Over Automation Procedures

Duolingo's Adaptive Learning: Space retrieval model, also known as spaced repetition, is a learning method in which information is revisited at progressively longer intervals to improve memory retention.

Research on spaced retrieval algorithms, such as the Leitner system [22] and SuperMemo's SM2 algorithm [23], has demonstrated the effectiveness of this method in reinforcing memory and preventing forgetting.

These algorithms are designed to prioritize the review of content just before it is likely to be forgotten, optimizing long-term retention by using data from previous learning interactions to schedule reviews more effectively. Studies in cognitive psychology, including the Ebbinghaus forgetting curve [24] provided a theoretical foundation that highlights how frequent, well-timed reviews lead to better recall and long-lasting learning outcomes.

Motivated by this body of research, Duolingo integrated these principles into its platform, initially, using a variant of the Leitner System (Fig. 4) to manage skill meter decay and practice schedule. The Leitner System uses a box system where each correct answer moves the learner's item to a box that increases the interval between reviews. This system allowed Duolingo to structure practice intervals effectively and ensure that users revisited content at optimal times to prevent forgetting [25].

However, to further personalized learning paths and boost efficiency, they developed their own model known as HalfLife Regression in 2016 [25]. This machine learning-based approach predicts when a learner is likely to forget a word or concept and schedules practice sessions accordingly. By applying the insights from spaced retrieval research, Duolingo has been able to optimize its lesson structure, tailoring reviews to each learner's individual performance and memory patterns, thus improving the effectiveness of its language-learning program.

Moreover, Henry, P. (2024) [26] elucidates how Duolingo employs Artificial Intelligence (AI) through a Large Language Model (LLM) to streamline lesson creation. This AI-powered LLM, known as "Birdbrain," predicts text completions and adjusts exercises' difficulty 'levels based on users' proficiency. To personalize practice, Duolingo had to conduct large-scale split testing (A/B testing) for each new Machine Learning (ML) model. This process doubled the amount of data processed and significantly increased computational demands.

To address this, Duolingo reduced the frequency of data writes to storage by buffering user data in memory and limiting database writes. This optimization minimized storage usage and reduced API latency, allowing Duolingo to test multiple models at half the cost required for testing a single model previously. These advancements also facilitated the introduction of new features, such as personalized vocabulary practice, and supported the continued evolution of its personalization systems [27].

Moreover, human experts collaborate with Birdbrain to ensure the content aligns with educational objectives while maximizing engagement [26]. By harnessing AI technology, Duolingo rapidly generates extensive lesson content with a single click, leading to enhanced efficiency, speed, and productivity in lesson development [26].

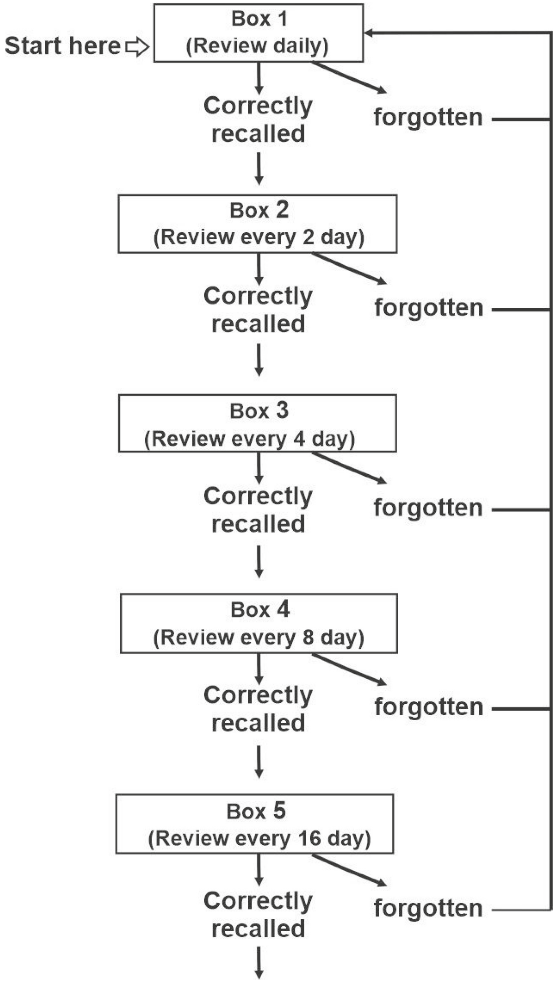


Fig. 4. Leitner System: You review flashcards from the box and test your recall. Depending on how well you recall the card, you either promote it to the next box or demote it back to Box 1 [25].

Duolingo's Communicative Approach: In embracing new pedagogical perspectives, Duolingo has shifted towards a communicative approach, departing from traditional grammar and translation methods. This evolution is exemplified by the introduction of innovative features like "Duolingo Stories" and Roleplay powered by GPT-4. These stories have undergone several changes since their inception. Initially, excluded from Duolingo's lesson structure, they are now seamlessly integrated [18].

Presented in two distinct difficulty levels, users initially engage with the content through reading and listening exercises. Most of the learner's time is spent reading and listening to the storylines with preset intervals for answering comprehension and vocabulary questions. The main focus of Duolingo Stories is therefore on improving reading and listening in terms of language skills and also vocabulary learning in terms of language components.

The improvement of the reading skill is limited to the genre of stories and/or fiction only. Even within this genre, the scope is further limited to dialogues between characters with the occasional use of third-person narration of the story. The listening skill is practiced in a relatively authentic manner. The voices are real human voices with careful consideration of suprasegmental features of phonology to elicit comprehension of the stories. The automatic replay of the audio at a slower pace, while helpful for some learners, further throws the authenticity of the task into question. Subsequent iterations introduced a challenge mode, offering a higher difficulty level that prioritizes listening skills, Here portions of the story are initially available only through auditory means, with textual support appearing upon user request [18].

Roleplay in Duolingo allows learners to practice real-world conversation skills through interactive, AI-driven scenarios with in-app characters. These exercises offer users the opportunity to engage in diverse conversations related to a real-life context. While not live interactions, the AI adapts dynamically, ensuring unique dialogues in each session.

After completing a conversation, learners receive AI-powered feedback on the accuracy and complexity of their responses, along with suggestions for improvement. Though AI facilitates their interactions, the underlying scenarios are crafted by Duolingo's human experts, ensuring that prompts align with the learner's progress. These experts also direct the flow of conversations to maintain relevance. Additionally, AI-generated responses are routinely reviewed for factual accuracy and tone to ensure a learning experience that is both educational and engaging [28].

The recent studies on Duolingo's effectiveness and its instructional methodologies highlight significant advancements in the platform's educational impact, showcasing its evolution from simple automation to sophisticated AI. For instance, a study quantifies Duolingo's efficiency, revealing that novice Spanish learners can achieve first-semester college proficiency with 26 to 49 hours of study, as indicated by significant gains in the WebCAPE placement test [29].

Furthermore, the study on L1-Spanish adult learners demonstrates that Duolingo's English course is particularly effective for developing general proficiency and receptive vocabulary, surpassing traditional classroom instruction in these areas. However, there is still room for improvement in listening skills and productive vocabulary [30].

A key illustration of the improvements stemming from the implementation of AI in the Duolingo application can be summarized in the following comparative Table 1.

Several improvements were gained by incorporating AI into traditional Automation. Both systems adjust based on interaction patterns, analyze user responses, and provide immediate feedback. However, AI enhances Duolingo by offering personalized feedback based on speech sound patterns, which is critical for pronunciation. Additionally, AI supports free-form writing suggestions and corrections adapting test proficiency levels, and including engaging characters with customizable voices and behaviors. These features significantly enhance user engagement and personalization, making the learning experience more effective and enjoyable [21].

Table 1. Comparative between Automation and Artificial Intelligence

Feature/ Functionality	Automation	Artificial Intelligence
Adjustment based on interaction patterns	x	x
Analysis of response to stimuli		x
Provision of immediate feedback	x	x
Provision of personalized feedback based on speech sound patterns		x
Writing exercise suggestions	x	x
Free-form writing exercise suggestions and corrections		x
Duolingo test proficiency adaptation		x
Inclusion of engaging characters	x	x
Customizable character voice and behavior		x

AI-Powered Video Call for Enhanced Speaking Practice: Additionally, a very recent feature added, Video call, aims to bridge the gap between structured language learning exercises and the demands of real-world conversational fluency [31].

This tool responds to a common challenge faced by language learners: the hesitation and difficulty in transitioning from understanding grammar and vocabulary to confidently engaging in everyday conversations. With Video Call, Duolingo provides a simulated environment where learners can develop their speaking skills without the pressure of real-time judgment or error correction.

This feature appears as an option at certain learning path nodes represented by a video camera icon. Learners can initiate a conversation with Lily, a virtual character who engages in real-time, AI-powered dialogues. This is the first time Duolingo has introduced a feature that allows learners to practice full conversation, as opposed to isolated phrases or sentences, thus promoting a more holistic approach to language practice.

The underlying AI models are capable of processing and generating natural language responses in real-time, a task that requires substantial computational power and linguistic modeling. These models are continually refined through machine learning algorithms that analyze user interactions, improving the accuracy and relevance of Lily's responses over time. For this reason, Video Call's outstanding features are its scalability and adaptability. Initially, conversations may last only one minute and involve simple phrases, but as learners advance, the length and complexity of the dialogue extend [31].

This gradual increase in difficulty aligns with Duolingo's broader adaptive learning model and its capacity for self-improvement ensures that the quality of conversations will continue to enhance as more users engage with the feature. Given the numerous advancements made possible by the integration of Artificial Intelligence into Duolingo's platform, a visual representation was developed to categorize and list these enhancements.

Figure 5 illustrates the various ways in which AI has contributed to improving both the pedagogical approach and computational efficiency of Duolingo's language learning platform. This representation is aimed at providing a clear framework for analyzing the impact of AI on the platform's educational outcomes.

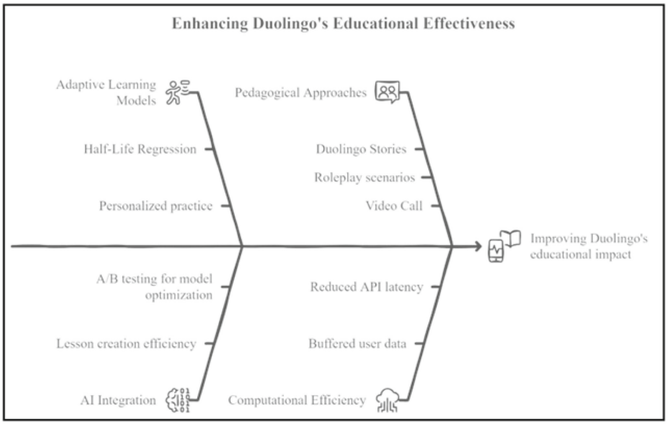


Fig. 5. Enhancing Duolingo's Educational Effectiveness powered by AI [18, 22, 25]–[27, 31]

6 User Experience Insights from Duolingo

To better understand how Duolingo is perceived as a language learning tool, a comparative analysis of user reviews from teachers, and young learners between the ages of 12 and 17 years old was conducted. The analysis involved 67 reviews from teachers and 54 reviews from younger learners, providing a comprehensive overview of the platform's reception across different demographic groups.

Teachers Review: Among the 67 reviews submitted by teachers expressed overwhelmingly positive experiences. Specifically, 49% of the reviews awarded Duolingo the highest possible rating of five stars, while 34% rated it four stars (see Fig. 6). This indicates that 83% of these users found the platform to be highly effective in either supplementing classroom instruction or facilitating personal language learning [31].

However, a smaller portion of users offered more moderate ratings, with 10 reviews (15%) giving three stars, and one (1%) rating Duolingo two stars. No one-star reviews were recorded in this group, reflecting a generally favorable perception of the platform among these users. Teachers specifically praised Duolingo's integration into the classroom as an engaging tool for reinforcing language learning. The gamified elements, ease of use, and flexibility in structuring lessons were highlighted as key strengths. On the other hand, some users expressed minor frustrations with the limited depth of the app's instructional content, noting that it might not be sufficient for advanced learners or for exclusive use without supplementary materials [31].

Young Learners Reviews: The 54 reviews from young learners (12-17 years old), who used Duolingo primarily for their own language learning, also leaned towards positive feedback. A notable 63% of young learners rated the platform five stars, while 24% gave it four stars (see Fig. 6). This suggests a high level of engagement and satisfaction among younger users. Additionally, only 11% provided a three-star rating, and 2% rated the app with two stars [32].

They often cited the app's interactive and game-like nature as the primary factor behind their enjoyment. The use of badges, streaks, and rewards appeared to motivate consistent use, while the simplicity of lessons made language learning accessible and fun. However, a few mentioned challenges in maintaining interest over time as the lessons progressed, echoing some of the concerns raised by older users about the app's ability to hold long-term engagement. Additionally, there were mentions of content featuring mature themes that may not be suitable for younger audiences [32].

Comparative Insights: When comparing both user groups, the key difference lies in the higher proportion of five-star ratings from young learners (63%) compared to teachers (49%). This may reflect Duolingo's effectiveness in appealing to younger audiences with its gamified approach, while older users might evaluate the app more critically considering the instructional depth alongside engagement. Despite these minor differences, both user groups largely agree on Duolingo's merits as a flexible and engaging language learning tool. The low number of two-and three-star reviews indicates a generally positive user experience across the board, with both groups recognizing the platform's accessibility and motivational features.

7 Limitation

While this paper provides a comprehensive analysis of Duolingo's evolution from automation to AI integration, several limitations must be acknowledged. First, the study primarily relied on secondary data, which may be constrained by the scope and biases of the original sources. This limitation could result in the exclusion of the most recent developments or proprietary data that are not publicly accessible.

Although the paper examines AI-driven tools such as ASR and Duolingo's Birdbrain AI, it does not empirically evaluate the effectiveness of these tools in real-world learning

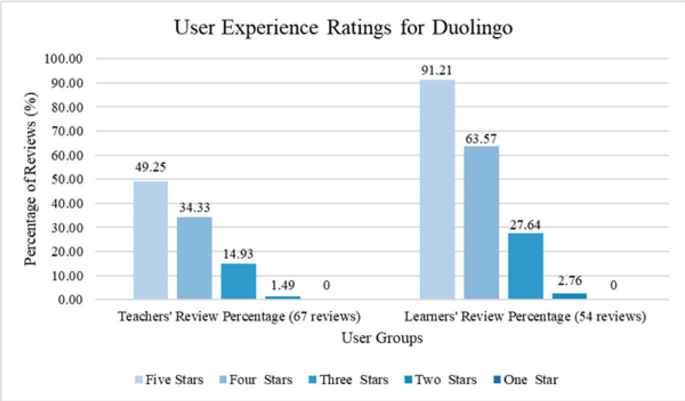


Fig. 6. Combined Duolingo Reviews Data from CommonSense and CommonSenseMedia [31, 32].

environments. This represents a gap in assessing their practical impact on language acquisition, Furthermore, while the focus on the migration from Python to Scala and the incorporation of AI is valuable, the paper does not explore the influence of more advanced AI models like GPT-3 and GPT-4, which could significantly affect learning outcomes.

Additionally, the analysis of user feedback, although insightful, presents certain limitations. User reviews are inherently subjective and may not accurately reflect the broader learner experience. Also, they often lack context regarding the user's background or language learning goals, and may be biased toward more engaged users who are more likely to leave feedback. This introduces potential bias in the conclusions drawn from the reviews, as they may not fully capture the range of user experiences.

Lastly, ethical considerations related to AI, such as data privacy, algorithmic bias, and the use of AI-driven content that may include mature themes inappropriate for children, were briefly touched upon but warrant a more thorough investigation. These issues are critical to ensuring that AI tools are both effective and ethically safe in the context of language learning.

8 Future Research

This paper opens several opportunities for future research. A promising direction is to investigate the learning efficacy of Duolingo's recently implemented AI functionalities through a comparative study of learning outcomes before and after AI integration.

Given that this study has primarily relied on secondary data, empirical studies focusing on specific user groups would provide a different understanding of AI's impact on language retention, fluency, and long-term learning outcomes. Additionally, further exploration is needed to address the ethical implications of AI integration in language learning platforms. This includes analyzing Duolingo's "Responsible AI practices" [21] and their potential effects on user privacy, data security, and algorithmic biases. Lastly, the evolution of Duolingo's technical stack, including its transition from the Scala programming language to more advanced generative AI models like GPT-3 and GPT-4, is a promising area for future inquiry. This exploration could examine how this shift impacts, both, the user experience and the platform's scalability in delivering personalized and adaptive learning.

9 Conclusions

1. Duolingo's journey from automation to Artificial Intelligence (AI) highlights significant advancements aimed at enhancing the user experience in language learning platforms. This paper has traced the historical progression of Computer-Assisted Language Learning (CALL), using Duolingo as a case study, demonstrating how technological and pedagogical innovations have shaped modern language platforms. CALL, initially rooted in behaviorist principles inspired by Skinner's work, has evolved toward more immersive and communicative approaches, in line with broader trends in language education.

The research questions examined the motivations behind Duolingo's transition from automated processes to the integration of AI tools, as well as the pedagogical shifts that accompanied this evolution. A secondary data analysis of Duolingo's development revealed key strategic decisions and milestones, including considerations of cognitive load in language learning and the ongoing expansion of the technological landscape.

2. Duolingo's automation, particularly through the integration of Automatic Speech Recognition (ASR) technology, marked a significant shift towards more interactive and personalized language learning experiences. ASR, often confused with AI, laid the groundwork for later AI-driven improvements by enabling error detection and providing accurate feedback. The introduction of AI-based tools like ASR into Duolingo's ecosystem revolutionized the learning process, offering users real-time feedback and personalized learning paths.

Moreover, Duolingo's shift from automation to AI has been driven by both sociological and technological factors, such as user demand for more advanced learning experiences and the need to address technical debt. The migration from Python to Scala marked a pivotal moment for Duolingo, enhancing the app's scalability, performance, and stability. Despite challenges like documentation gaps and compatibility issues, these technological decisions significantly improved both user experience and learning outcomes.

3. This study opens several opportunities for future re-search. One promising direction is to evaluate the learning effectiveness of the Duolingo app, specifically by comparing learning outcomes before and after the implementation of AI. Additionally, further research could explore the ethical implications of AI integration, examining Duolingo's "Responsible AI Practices" and their impact on user privacy. Lastly, an exploration of Duolingo's transition toward generative AI models such as GPT-3 and GPT-4 could offer further insights, as these models were not covered in this study.

4. In conclusion, Duolingo embodies broader trends at the intersection of technology and language education. As language learning methods and AI capabilities continue to evolve, Duolingo's approach serves as a model for future educational platforms that seek to blend advanced technology with effective pedagogical practices. Through its commitment to innovation, Duolingo demonstrates the potential for AI to create more personalized, accessible, and impactful learning experiences for users worldwide.

References

1. Anderies, C., Agustina, T., Lipiena, A., Raaziqi, A., Gunawan, A.A.S.: User experience analysis of Duolingo using user experience questionnaire. Eng. Math. Comput. Sci. J. (EMACS) **5**(3), 155–159 (2023). https://doi.org/10.21512/emacsjournal.v5i3.9227
2. Horie, A.K.: Rewriting duolingo's engine in scala, Duolingo Blog. https://blog.duolingo.com/rewriting-duolingos-engine-in-scala/ (2020)
3. Chapelle, C.A., Sauro, S.: The Handbook of Technology and Second Language Teaching and Learning. Wiley (2017)
4. Kumar, E.S., Sreehari, P.: Computer assisted language learning (call). In: Foundation Books, pp. 3–14 (2009). https://www.cambridge.org/core/books/abs/handbook-for-english-language-laboratories/computer-assisted-language-learning-call/97FCB7A84A31665A92E7C3D2051D0D54, https://doi.org/10.1017/upo9788175968677.003
5. Zeng, S.: The potential of online technology for language learning. English Lang. Teach. **13**(10), 23 (2020). https://doi.org/10.5539/elt.v13n10p23, https://doi.org/10.5539/elt.v13n10p23
6. Richards, J.C.: Communicative Language Teaching Today (2006)
7. Rintaningrum, R.: Technology integration in English language teaching and learning: Benefits and challenges. Cogent Educ. **10**(1) (2023). https://doi.org/10.1080/2331186x.2022.2164690
8. Segovia, C., Sigcha, E.A.P., Alarco´n, C.A.A.: Duolingo app in the development of oral comprehension. Explorador Digital **7**(4), 19– 33 (2023). https://doi.org/10.33262/exploradordigital.v7i4.2729
9. Adams, S.: Game of tongues: how duolingo built a $700 million business with its addictive language-learning app. Forbes (2019). https://www.forbes.com/sites/susanadams/2019/07/16/game-of-tongues-how-duolingo-built-a-700-million-business-with-its-addictive-language-learning-app/?sh=2b6158b63463
10. GoodWhale: Language learning made free: The fascinating journey of Duolingo's founder, Luis von ahn (2023). https://www.linkedin.com/pulse/language-learning-made-free-fascinating-journey-duolingos
11. Mayer, R.E., Fiorella, L., (eds.) The Cambridge Handbook of Multimedia Learning. Cambridge University Press (2021). http://www.cambridge.org/9781107035201
12. Medical College of Wisconsin: Cognitive load theory: a guide to applying cognitive load theory to your teaching, Medical College of Wisconsin (2022). https://www.mcw.edu/-/media/MCW/Education/Academic-Affairs/OEI/Faculty-Quick-Guides/Cognitive-Load-Theory.pdf
13. Bahari, A., Wu, S., Ayres, P.: Improving computer-assisted language learning through the lens of cognitive load. Educ. Psychol. Rev. **35**(2) (2023). https://doi.org/10.1007/s10648-023-09764-y
14. Neri, A., Cucchiarini, C., Strik, H.: Automatic speech recognition for second language learning: How and why it actually works. In: Proceedings of the 2003 International Conference on Speech Communication and Technology, pp. 1157–1160 (2003). http://lands.let.ru.nl/literature/neri.2003.1.pdf,

15. IBM, ¿que´ es el reconocimiento del habla? IBM. https://www.ibm.com/es-es/topics/speech-recognition
16. Jones, D.M., Cheng, L., Tweedie, M.G.: Automated scoring of speaking and writing: Starting to hit its stride. Can. J. Learn. Technol. **48**(3), 1–22 (2023). https://doi.org/10.21432/cjlt28241
17. Gekhman, Z., Zverinski, D., Mallinson, J., Beryozkin, G.: Red-ace: robust error detection for asr using confidence embeddings, arXiv preprint. arXiv:2203.07172 (2022). https://doi.org/10.48550/arxiv.2203.07172
18. Yousefpoori-Naeim, M.: A review of duolingo stories. Int. J. Innov. Online Educ. (2023). https://doi.org/10.1615/IntJInnovOnlineEdu.2023048899
19. What is python? Executive summary, Python.org. https://www.python.org/doc/essays/blurb/?external_link=true
20. Tour of scala, Scala documentation. https://docs.scala-lang.org/tour/tour-of-scala.html
21. Wodzak, S.: How duolingo is using artificial intelligence for social good, Duolingo Blog (2024). https://blog.duolingo.com/ai-improves-education/
22. Leitner, S.: So Lernt Man Lernen. Herder Verlag (1972)
23. Wozniak, P.A.: SuperMemo: Optimization of Learning (1990)
24. Ebbinghaus, H.: Memory: A Contribution to Experimental Psychology. Dover Publications (1885)
25. Settles, B., Meeder, B.: A trainable spaced repetition model for language learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/p16-1174
26. Henry, P.: How duolingo uses ai to create lessons faster, Duolingo Blog (2024). https://blog.duolingo.com/large-language-model-duolingo-lessons/
27. Team, D.: Inside engineering at duolingo, Duolingo Blog (2024). https://blog.duolingo.com/unique-engineering-problems/
28. Team, D.: Duolingo max uses openai's gpt-4 for new learning features, Duolingo Blog (2024). https://blog.duolingo.com/duolingo-max/
29. Grego, J.: Duolingo effectiveness study final report (2012). https://www.languagezen.com/pt/about/english/Duolingo_Efficacy_Study.pdf
30. Gonza´lez-Ferna´ndez, B.: The effectiveness of duolingo vs. classroom instruction on Spanish speakers' l2 english proficiency and lexical development (2023). https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_language_read_listen_write_speak_2024.pdf
31. Team, D.: Duolingo's video call offers realistic conver- sation practice. Duolingo Blog (2024). https://blog.duolingo.com/video-call/
32. Duolingo review for teachers, Common Sense Edu-cation, https://www.commonsense.org/education/reviews/duolingo

Hybrid SCA-IPOPT Approach for the Optimal Location and Sizing of TSCs in Medium-Voltage Distribution Networks

Oscar Danilo Montoya[1], Víctor M. Garrido-Arévalo[2](✉), and Walter Gil-González[3]

[1] Grupo de Compatibilidad e Interferencia Electromagnética (GCEM), Facultad de Ingeniería, Universidad Distrital Fracisco José de Caldas, Bogotá D.C. 110121, Colombia
odmontoyag@udistrital.edu.co

[2] Programa de Ingeniería Eléctrica, Universidad de Pamplona, Pamplona 760042, Colombia
victor.garrido@unipamplona.edu.co

[3] Departamento de Ingneiería Eléctrica, Facultad de Ingenierías, Universidad Tecnológica de Pereira, Pereira 660003, Colombia
wjgil@utp.edu.co

Abstract. This paper explores the optimal siting and sizing of thyristor-switched capacitors (TSCs) in distribution networks, aiming to minimize the operating costs by reducing energy losses while accounting for the annualized TSC investment costs. Within the optimization strategy, the sine-cosine algorithm (SCA) determines the optimal TSC locations and sizes, and the interior-point optimizer (IPOPT) solves the resulting optimal power flow problem, ensuring an efficient grid operation under fixed and variable reactive power scenarios. Tested on a 33-bus system, the proposed methodology was compared against the Chu & Beasley genetic algorithm (CBGA), the particle swarm optimizer (PSO), the black widow optimizer (BWO), and the SCA. According to the results, the BWO, SCA, and SCA-IPOPT methodologies reached the best solutions, reducing the objective function by 11.22%, with SCA-IPOPT further improving this value to 12.43% under a variable reactive power scenario. All numerical validations were performed using the Julia software.

Keywords: thyristor-switched capacitors · sine-cosine algorithm · interior-point optimization · fixed and variable reactive power injection · optimal power flow solution · hybrid optimization methodology

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 72–88, 2025.
https://doi.org/10.1007/978-3-031-88854-0_6

1 Introduction

1.1 General Context

Reactive power compensation has long been a key strategy for minimizing power losses and improving voltage profiles in electrical distribution networks with resistive-inductive loads [20]. Traditionally, this has been achieved by integrating capacitor-based devices to inject reactive power according to demand profiles [18]. Conventional methods utilize fixed or variable-step capacitor banks strategically placed across the grid. However, while effective, these systems fall short in addressing dynamic fluctuations in the active and reactive power demand [15]. Fixed reactive power injections often lack the flexibility to adapt to real-time changes in grid conditions, which leads to inefficiencies. To address this limitation, more advanced solutions are required, such as flexible alternating current transmission systems (FACTS) [20]. Technologies like static VAR compensators (SVCs), static synchronous compensators (STATCOMs), and thyristor-switched capacitors (TSCs) offer a more precise and responsive approach, enabling the real-time adjustment of reactive power injections and enhancing the overall efficiency and stability of the network [4].

2 Motivation

In certain applications, TSCs have several distinct advantages over other reactive power compensation technologies (*e.g.*, SVCs and STATCOMs) [9]. One of the main benefits is their lower initial capital cost. Given their simpler design and the absence of advanced power electronic components, TSCs offer a more economical solution for medium-voltage distribution networks, particularly in scenarios where budget constraints are a significant factor. This makes them an attractive alternative for distribution companies seeking to minimize upfront investments while still enhancing grid performance [13].

In addition to cost-effectiveness, TSCs can be more efficient than SVCs and STATCOMs in terms of energy losses, especially in applications that demand reactive power compensation. Their ability to provide reactive power with fewer energy losses contributes to a better overall system efficiency, which is critical in optimizing the performance of electrical networks. Furthermore, TSCs have a longer lifespan and lower maintenance requirements compared to their more complex counterparts. The simplified design of TSCs reduces the likelihood of component failures, leading to increased reliability and decreased maintenance costs over time [?].

Given these benefits, there is a strong motivation to explore the optimal integration of TSCs in modern distribution networks, particularly in the context of increasing energy demands and the need for more flexible and cost-effective reactive power management solutions. This research aims to study the potential of TSCs to enhance grid stability and efficiency while offering a more sustainable and economical alternative to traditional FACTS [13].

2.1 Literature Review

The specialized literature includes several proposals for the optimal integration (siting and sizing) of FACTS in electrical networks. Some of these works are presented below, in order to outline the main developments in the field of power systems.

The paper proposed by [1] used the sine-cosine algorithm (SCA) to solve the problem regarding the optimal integration of FACTS and VAR sources in power systems. The proposed algorithm was implemented on the IEEE 57-bus test system to evaluate and compare its performance against many optimization approaches like particle swarm optimization and the crow search algorithm, among others. This work indicates that, among the compared methods, SCA exhibits the best performance in minimizing system losses and finding the optimal placement of FACTS, which demonstrates its effectiveness in reactive power planning.

In this context, a recent study discussed the integration of FACTS to enhance the stability of power systems [17]. The authors used SVCs and TSCs on the IEEE 39-bus system to improve voltage stability and power flow control. A PSO-based technique was applied to determine the optimal location and size of FACTS, in order to achieve a large system loading capacity or performance improvements. This feasibility study pointed out that FACTS may provide a solution to the problems associated with the increasing adoption of electric vehicles and that, in general, they make power networks more stable and efficient.

The authors of [8] applied the seeker optimization algorithm (SOA) to the optimal integration of TSCs into transmission systems. They used the IEEE 30-bus system and an Iraqi 400 kV network to demonstrate how TSCs can improve voltage profiles, reduce power losses, and increase the capacity of transmission lines. The SOA outperformed the traditional approaches in efficiently minimizing costs while enhancing the stability of TSCs.

Several methods have been proposed to optimally integrate FACTS in distribution systems. For instance, in [22], a novel hybrid method was proposed which combined fuzzy multi-objective optimization with ant colony algorithms. It was developed to optimize system reconfigurations involving the placement of FACTS alongside distributed solar energy generators, aiming to reduce energy losses and improve voltage profiles.

In [7], a heuristic algorithm focusing on power losses and voltage indices was proposed in order to determine the optimal integration of FACTS into electrical distribution systems. The heuristic algorithm was evaluated on the IEEE 33-bus test feeder under peak conditions, yielding good results.

The work by [21] proposed an analytical method hybridized with heuristic optimization techniques for the optimal location and sizing of FACTS. This study also mentions some of the most commonly applied objective functions in this field, such as the improvement of voltage stability and the minimization of power losses.

Using the multi-objective PSO algorithm, the authors of [19] determined the optimal location and size of FACTS while considering network reconfiguration.

The proposed multi-objective method minimized the active energy losses while increasing voltage stability indices and decreasing the load capacity factor of the transmission lines. However, a drawback of this proposal is that it only considered maximum load conditions.

It should be pointed out that most of the available literature does not consider the deployment aspects of TSCs, which may be superior in terms of performance to other alternatives such as SVCs and D-STATCOMs.

2.2 Contributions and Scope

In light of the above-presented literature review, the main contribution of this research is the development of a hybrid optimization methodology that combines the SCA with the interior-point optimizer (IPOPT) in order to determine the optimal location and size of TSCs in electrical distribution networks while minimizing the expected annual operating costs. Numerical simulations conducted in the 33-bus system demonstrate that allowing TSCs to dynamically adjust their reactive power injection over time leads to significant reductions in the objective function value when compared to a fixed-injection strategy.

Remark 1. *In a previous report by [13], the authors applied the SCA approach for the siting and sizing of TSCs in distribution networks via a master-slave optimization framework. In this contribution, the slave stage is enhanced by replacing classical power flow theory with the IPOPT approach. This modification allows determining the operation of TSCs throughout the day, considering either fixed or variable reactive power injections. The key advantage of the IPOPT is that it ensures a global optimal solution for the nonlinear programming problem associated with each combination of TSC sizes and locations provided by the SCA.*

2.3 Document Structure

The remainder of this document is organized as follows. Section 3 outlines the general formulation for siting and sizing TSCs in electrical distribution networks, with the goal of minimizing the expected annual operating costs associated with energy losses while accounting for the investment costs of reactive power compensators. Section 4 describes the solution methodology, which involves the hybridization of the SCA and IPOPT. The SCA is employed to determine the optimal locations and sizes of the TSCs, while the IPOPT is entrusted with defining the daily optimal dispatch of these devices in the context of the optimal power flow (OPF) problem. Section 5 details the characteristics of the test system used for validation (*i.e.*, a 33-bus grid), incorporating daily variations in the active and reactive power curves as well as the parameterization of the investment costs related to the TSCs. Section 6 presents the numerical results obtained and an analysis of the proposed hybrid SCA-IPOPT approach that considered both fixed and variable reactive power injections. Our approach is also compared against several metaheuristic optimizers from the literature. Finally, Sect. 7 provides the key concluding remarks of this work and suggests potential directions for future research.

3 Mathematical Modeling

The mathematical model for the optimal integration of TSCs into electrical distribution networks is classified as a mixed-integer nonlinear programming (MINLP) problem. Here, the integer variables represent the selection of nodes for the installation of reactive power compensators, while the continuous variables correspond to the solution of the power flow equations [20].

3.1 Objective Function

This research focuses on the integration of TSCs in distribution grids, with the objective of minimizing the annual operating costs denoted by f [13]. These costs include both the energy losses and the installation expenses associated with TSCs. The optimization process is governed by the objective function outlined in (1).

$$\min\ f = f_1 + f_2, \tag{1}$$

$$f_1 = CT \sum_{h \in \mathcal{H}} \sum_{l \in \mathcal{B}} \left(p_{lh}^{s} + p_{lh}^{r}\right) \Delta h, \tag{2}$$

$$f_2 = T\left(\frac{k_1}{k_2}\right) \sum_{k \in \mathcal{N}} q_k^{\mathrm{TSC}} \left(\omega_1 \left(q_k^{\mathrm{TSC}}\right)^2 + \omega_2 q_k^{\mathrm{TSC}} + \omega_3\right), \tag{3}$$

In the given expression, f_1 represents the annual costs associated with energy losses in the system, while f_2 corresponds to the investment costs related to the installation of TSCs. The parameters C and T refer to the average cost of energy losses and the number of days in a year, respectively. The variables p_{lh}^{s} and p_{lh}^{r} denote the active power flows sent and received by a distribution line, respectively. The term Δh represents the time interval used in the daily analysis (0.5 h), and q_k^{TSC} specifies the nominal capacity of the TSCs. The coefficients ω_1, ω_2, and ω_3 are polynomial terms of the objective function f_2, which represent the investment costs of the TSCs. Additionally, the constants $k_1 > 0$ and $k_2 > 0$ account for the annualized investment costs, distributed over a ten-year planning horizon [13].

3.2 Model Constraints

The constraints associated with the problem under study involve several key factors [13]. These include the power balance equations for each node concerning both active and reactive power, limits on the maximum and minimum power flow through the distribution lines, voltage regulation bounds, and restrictions on the number of TSC devices to be installed [20]. Additionally, the TSCs' capacity to inject or absorb reactive power must be considered, along with other relevant operating constraints.

3.2.1 Power Equilibrium per Node

The nodal power balance equations for node k, in relation to both active and reactive power, account for the sums of generated and demanded power. These sums are equated to the injected active or reactive power flows [20]. Using the branch power flow formulation, the equations can be expressed as follows [14]:

$$p_{kh}^{g} - p_{kh}^{d} = \sum_{l \in L} \left(A_{kl}^{+} p_{lh}^{s} + A_{kl}^{-} p_{lh}^{r} \right), \forall\ k \in \mathcal{N}, \forall\ h \in \mathcal{H} \tag{4}$$

$$q_{kh}^{g} - q_{kh}^{d} + q_{kh}^{\mathrm{TSC}} = \sum_{l \in L} \left(A_{kl}^{+} q_{lh}^{s} + A_{kl}^{-} q_{lh}^{r} \right), \forall\ k \in \mathcal{N}, \forall\ h \in \mathcal{H} \tag{5}$$

where p_{kh}^{g} and p_{kh}^{d} represent the generated and demanded active power at node k, respectively; q_{kh}^{g} and q_{kh}^{d} denote the generated and demanded reactive power at the same node; the terms p_{lh}^{s} and q_{lh}^{s} represent the active and reactive power flows sent through distribution line l; p_{lh}^{r} and q_{lh}^{r} correspond to the respective receiving power flows; the variable q_{kh}^{TSC} denotes the reactive power injected by the TSCs at node k and time h; and A^{+} and A^{-} are the positive and negative entries of the node-to-branch incidence matrix A, where $A = A^{+} + A^{-}$.

3.2.2 Branch Power Flow Calculations

The active and reactive power flowing through a given distribution line, along with its corresponding maximum capacities, is expressed as follows [6]:

$$p_{lh}^{s} = \mathrm{real}\left(v_{kh} y_{l}^{*} \left(v_{kh} - v_{mh} \right)^{*} \right), \forall\ l \in \mathcal{B},\ \forall\ h \in \mathcal{H}, \tag{6}$$

$$p_{lh}^{r} = \mathrm{real}\left(v_{mh} y_{l}^{*} \left(v_{mh} - v_{kh} \right)^{*} \right), \forall\ l \in \mathcal{B},\ \forall\ h \in \mathcal{H}, \tag{7}$$

$$q_{lh}^{s} = \mathrm{imag}\left(v_{kh} y_{l}^{*} \left(v_{kh} - v_{mh} \right)^{*} \right), \forall\ l \in \mathcal{B},\ \forall\ h \in \mathcal{H}, \tag{8}$$

$$q_{lh}^{r} = \mathrm{imag}\left(v_{mh} y_{l}^{*} \left(v_{mh} - v_{kh} \right)^{*} \right), \forall\ l \in \mathcal{B},\ \forall\ h \in \mathcal{H}, \tag{9}$$

$$\| p_{lh}^{s} + j q_{lh}^{s} \| \leq s_{l}^{\max},\ \forall\ l \in \mathcal{B},\ \forall\ h \in \mathcal{H} \tag{10}$$

$$\| p_{lh}^{r} + j q_{lh}^{r} \| \leq s_{l}^{\max},\ \forall\ l \in \mathcal{B},\ \forall\ h \in \mathcal{H}. \tag{11}$$

In this formulation, v_{kh} and v_{mh} denote the complex voltages at nodes k and m, respectively, while y_l represents the admittance of branch l. The term $s_{l}^{\max}$ denotes the maximum allowable apparent power that can flow through the line.

3.2.3 Operating Conditions of the Distribution Network

The voltage levels at each node of an electrical network must remain within the limits set by regulatory standards, thereby ensuring the network's proper and safe operation [11]. These limits include:

$$v_{0h} = v^{\mathrm{nom}} e^{j0}, \forall\ h \in \mathcal{H}, \tag{12}$$

$$\| v_{kh} \| \geq v^{\min}, \forall\ k \in \mathcal{N},\ \forall\ h \in \mathcal{H}, \tag{13}$$

$$\| v_{kh} \| \leq v^{\max}, \forall\ k \in \mathcal{N},\ \forall\ h \in \mathcal{H}. \tag{14}$$

Here, v_{0h} denotes the nominal voltage at the slack node (substation) at time h, with v^{nom} referring to its typical value, usually set to 1.0 per unit. Additionally, v^{min} and v^{max} specify the allowable voltage range, indicating the minimum and maximum levels permitted within the distribution grid.

3.2.4 Siting and Sizing the TSCs

The optimal integration of TSCs requires addressing two key factors: identifying their ideal locations within an electrical grid and determining their optimal size [20]. To tackle these challenges, the following constraints are defined:

$$0 \leq q_k^{\text{TSC}} \leq z_k q_{\text{max}}^{\text{TSC}},\ \forall\ k \in \mathcal{N}, \tag{15}$$

$$-q_k^{\text{TSC}} \leq q_{kh}^{\text{TSC}} \leq q_k^{\text{TSC}},\ \forall\ k \in \mathcal{N},\ \forall\ h \in \mathcal{H} \tag{16}$$

$$\sum_{k \in \mathcal{N}} z_k \leq \eta, \tag{17}$$

$$z_k \in \{0, 1\},\ \forall\ k \in \mathcal{N}. \tag{18}$$

In this formulation, q_k^{TSC} is a variable that defines the size of the TSCs to be installed, with $q_{\text{max}}^{\text{TSC}}$ indicating their maximum capacity. This study assumes TSC capacities ranging from 0 Mvar to 2 Mvar, a typical range for distribution networks [13]. The binary vector z indicates the presence of TSCs at node k, where $z = 1$ signifies installation and $z = 0$ indicates absence. Furthermore, η denotes the maximum number of TSCs allowed for installation.

4 Solution Methodology

In this research, the optimal siting and sizing of TSCs is achieved through a novel hybrid optimization methodology that combines the strengths of the SCA [13] with the IPOPT. The former is entrusted with identifying the optimal location of the TSCs and estimating their sizes, while the latter is used to solve the equivalent branch power flow problem, determining both the operating costs and the daily power injection profiles of the grid. The key details of the SCA and the IPOPT are presented below.

4.1 The Sine-Cosine Algorithm

The SCA is a mathematics-inspired metaheuristic optimization method that uses trigonometric functions to explore the solution space [5]. It is well-suited for siting and sizing TSCs, as it effectively handles mixed discrete-continuous optimization problems. Initially proposed in [12], this algorithm was developed to tackle large-scale nonlinear optimization problems such as the optimal power flow in transmission systems.

This research employs the SCA to determine variables such as the optimal locations (discrete component) and the sizes (continuous component) of the TSCs. The key steps of the algorithm are outlined below.

4.1.1 Problem Encoding and Initial Population

For the problem under study, the proposed encoding to represent the location and sizes of the TCSs is shown in Eq. (19).

$$x_j^p = \underbrace{z_1\ z_2 \cdots z_n}_{\text{Location}} \quad \underbrace{q_1^{\text{TSC}}\ q_2^{\text{TSC}} \cdots q_n^{\text{TSC}}}_{\text{Sizing}}. \tag{19}$$

In Eq. (19), x_j represents the j^{th} candidate solution within the population at iteration p (denoted as X^p). Here, X^p is a matrix with dimensions $n_s \times n_v$, where n_s is the number of candidate solutions and n_v the number of decision variables.

It is important to note that, to generate each x_j^p (with $p = 0$), the generation rule in Eq. (20) is applied:

$$x_{j,l} = x_l^{\min} + \beta_l \left(x_l^{\max} - x_l^{\min} \right), \begin{Bmatrix} l = 1, 2, \ldots, n_v \\ j = 1, 2, \ldots, n_s \end{Bmatrix} \tag{20}$$

where β_l is a random variable uniformly distributed between 0 and 1, and $x_l^{\min}$ and $x_l^{\max}$ represent the lower and upper bounds for the decision variables, respectively. Also note that the first part of the vector corresponds to discrete variables, indicating the nodes where the TCSs will be installed. Therefore, this component must be rounded to the nearest integer.

4.1.2 Evolution Rules

Once the initial population X^p has been generated at $p = 0$, the next step is to identify the optimal solution within that population, denoted as x_{best}^p. To determine the optimal solution at iteration p, each of the n_s candidate solutions in the population X^p must be evaluated using the problem's objective function (this will be the solution obtained by the IPOPT). This evaluation allows selecting the best-performing solution:

$$x_{\text{best}}^p = \left\{ x_j^p / x_j^p \rightarrow \min \left(F_f \left(x_j^t \right) \right), \ \forall j = 1, 2, \ldots, n_s \right\}, \tag{21}$$

where x_j^p represents the j^{th} solution at iteration p, and $F_f \left(x_j^t \right)$ denotes the objective function evaluated with the solution x_j^t.

Once the current best solution has been identified, the update rule described in Eq. (25) is applied for all $k = 1, 2, \ldots, n_s$:

$$y_j^t = \begin{cases} x_j^t + r_1 \sin(r_2) \cdot \left| r_3 x_{\text{best}}^p - x_j^t \right| & r_4 \le \frac{1}{2} \\ x_j^t + r_1 \cos(r_2) \cdot \left| r_3 x_{\text{best}}^p - x_j^t \right| & r_4 > \frac{1}{2} \end{cases}, \tag{22}$$

where y_j^t represents the potential new solution, and r_1 is the linear decay function. The vector r_2 contains random variables uniformly distributed in the range $[0, 2\pi]$, with dimensions $1 \times n_v$, while r_3 and r_4 are random values between 0 and 1, obtained from a uniform distribution. The notation $v \cdot w$ indicates an element-wise multiplication between the vectors v and w. Additionally, r_1 is recalculated at each iteration.

$$r_1 = a\left(1 - p/p_{\max}\right). \tag{23}$$

In this context, $p_{\max}$ represents the maximum number of iterations, and a is a scale factor chosen by the programmer. According to [12], this scalar is commonly set to a value of 2 ($a = 2$).

Each potential solution y_k^t must be evaluated to ensure compliance with the constraints imposed on the decision variables. Accordingly, the correction rule described in Eq. (24) is applied to each variable for all $j = 1, 2, \ldots, n_v$ and $l = 1, 2, \ldots, n_s$:

$$y_{j,l}^p = \begin{cases} y_{j,l} & x_l^{\min} \leq y_{j,l} \leq x_l^{\max} \\ x_l^{\min} & y_{j,l} < x_l^{\min} \\ x_l^{\max} & y_{j,l} > x_l^{\max} \end{cases}. \tag{24}$$

It should be noted that the discrete component of the optimization problem must be rounded to ensure the feasibility of each solution before evaluating it in the fitness function.

4.1.3 Population Substitution

Once the new potential solutions (y_k^t) have been generated, it is essential to determine whether they should be included in the population for the next iteration. This decision is based on an evaluation of the performance function $F_f\left(y_j^t\right)$. The population update rule, as outlined in Eq. (??), is used to decide which solutions are retained:

$$x_j^{p+1} = \begin{cases} y_j^p, \ F_f\left(y_j^p\right) < f\left(x_j^p\right) \\ x_j^p, \ F_f\left(y_j^p\right) \geq f\left(x_j^p\right) \end{cases}, \quad \{\forall j = 1, 2, \ldots, n_s\}. \tag{25}$$

4.1.4 End of the Solution Process

When applying metaheuristic optimization techniques, two common stopping criteria are frequently referenced in the literature. The process of exploring and exploiting the solution space can be terminated (i) when the maximum number of iterations set by the programmer is reached, or (ii) when no significant improvement in the performance function is observed after $k_{\max}$ consecutive iterations. The latter requires a counter to monitor improvements in the objective function. If this criterion is used, the value of $k_{\max}$ is defined by the user, with a suggested value between 10% and 30% of the total iterations.

4.2 The Interior-Point Optimizer

The IPOPT is a powerful numerical method used to solve large-scale nonlinear optimization problems. It operates by iteratively moving through the interior of the feasible region, seeking optimal solutions while satisfying the established constraints [23]. Unlike boundary-focused methods, the IPOPT efficiently explores the solution space by maintaining feasibility concerning the inequality constraints throughout the process. This approach is particularly advantageous in nonlinear optimization due to its ability to handle complex non-convex problems with both equality and inequality constraints [16]. Its robustness, scalability, and quick convergence make it an ideal choice for large-scale engineering and industrial applications where precision and efficiency are critical. Additionally, the IPOPT's effectiveness in handling smooth nonlinear objective functions and constraints makes it well-suited for applications in science and engineering [2].

The general formulation of an optimization problem that can be solved using the IPOPT has the following structure:

$$\text{Minimize} \quad f_0(x), \tag{26}$$

$$\text{s.t.:} \quad f(x) \sim b_1, \tag{27}$$

$$A_L x \sim b_2, \tag{28}$$

$$l_b \le x \le u_b, \tag{29}$$

where $\sim$ defines the vector of relational operations (*i.e.*, $\le$, $=$, or $\ge$); $x \in \mathbb{R}^n$ is the n-dimensional vector of decision variables; $f_0(x) : \mathbb{R}^n \to \mathbb{R}$ is a nonlinear soft function associated with the objective function to be minimized; $l_b \in \mathbb{R}^n$ and $b_b \in \mathbb{R}^n$ denote the lower and upper limits of the decision variable vector; $f(x) : \mathbb{R}^n \to \mathbb{R}^{m_1}$ are m_1 soft nonlinear functions associated with the non-affine constraints, with $b_2 \in \mathbb{R}^{m_2}$ being a constant vector related to their right-hand-side limits; $A_L \in \mathbb{R}^{m_2 \times n}$ is a sparse matrix associated with the linear (*i.e.*, affine) constraints; and $b_2 \in \mathbb{R}^{m_2}$ is a constant vector associated with the parameters of the affine constraints.

Remark 2. *It is important to note that, in the Optimization Model (1)–(18), once the variables related to the location and size of the TSCs (z_k and q_k^{TSC}) have been determined by the SCA, the model is reduced to an equivalent nonlinear optimization (NLP) problem, as defined in (26)–(29). This transformation allows efficiently solving the problem via the IPOPT, which yields the objective function value and enables the SCA's evolution throughout the solution space.*

4.3 Summary of the Optimization Methodology

Figure 1 presents the flowchart used for the computational implementation of the proposed optimization approach.

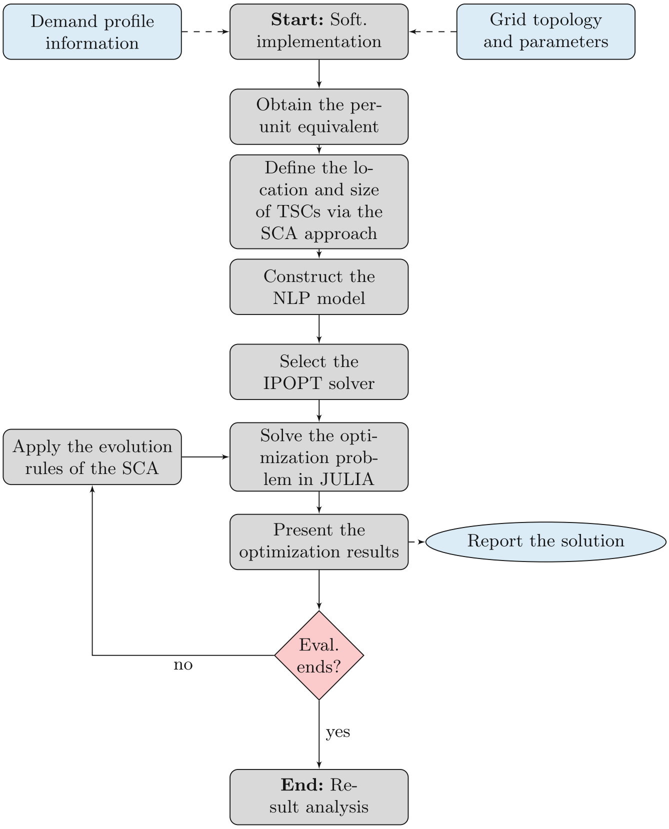


Fig. 1. Solution methodology combining the SCA and IPOPT approaches

Note that our selection of the IPOPT for solving the studied problem was based on its effectiveness in addressing NLP problems [15], as well as on literature reports regarding power system applications [?], optimal power flow solutions [3], and electric distribution networks [24], among others.

5 Test Feeder

To validate the effectiveness of the proposed optimization approach for locating and sizing TSCs in electrical distribution networks, a 33-bus grid was employed. The electrical configuration of this test feeder and all the numerical information regarding its branches and power consumption are presented in Fig. 2 and Table 1. These data were obtained from [20].

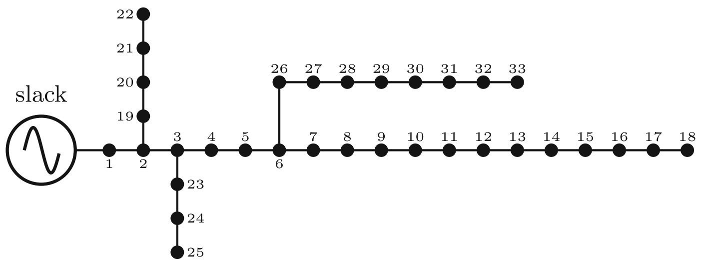


Fig. 2. Electrical configuration of a 33-bus grid system used

Typically, electrical distribution systems undergo daily load fluctuations. The demand curves regarding both active and reactive power are presented in Fig. 3. These curves exemplify the typical behavior observed in a Colombian electrical distribution system [20].

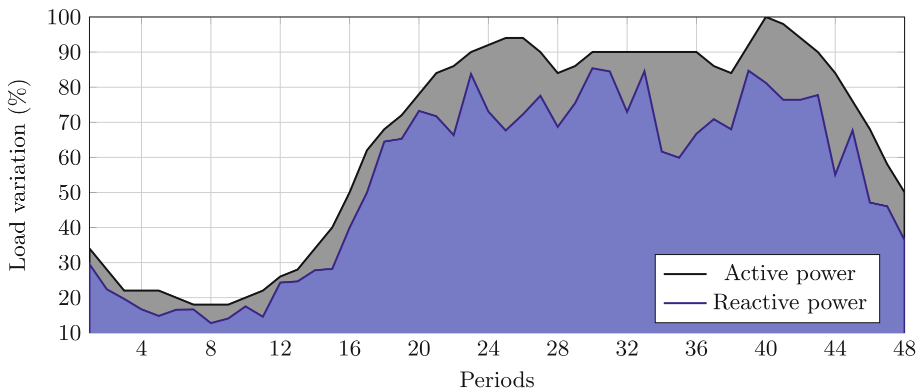


Fig. 3. Daily expected variations regarding active and reactive power consumption

To evaluate the Objective Function (1) and its components f_1 and f_2 related to the optimal siting and sizing of TSCs, the parameters listed in Table 2 were considered [20].

Table 1. Electrical parameters of the 33-test system used

k	m	R_{km} (Ω)	x_{km} (Ω)	P_k (kW)	Q_k (kW)
1	2	0.0922	0.0477	100	60
2	3	0.4930	0.2511	90	40
3	4	0.3660	0.1864	120	80
4	5	0.3811	0.1941	60	30
5	6	0.8190	0.7070	60	20
6	7	0.1872	0.6188	200	100
7	8	17.114	12.351	200	100
8	9	10.300	0.7400	60	20
9	10	10.400	0.7400	60	20
10	11	0.1966	0.0650	45	30
11	12	0.3744	0.1238	60	35
12	13	14.680	11.550	60	35
13	14	0.5416	0.7129	120	80
14	15	0.5910	0.5260	60	10
15	16	0.7463	0.5450	60	20
16	17	12.890	17.210	60	20
17	18	0.7320	0.5740	90	40
2	19	0.1640	0.1565	90	40
19	20	15.042	13.554	90	40
20	21	0.4095	0.4784	90	40
21	22	0.7089	0.9373	90	40
3	23	0.4512	0.3083	90	50
23	24	0.8980	0.7091	420	200
24	25	0.8960	0.7011	420	200
6	26	0.2030	0.1034	60	25
26	27	0.2842	0.1447	60	25
27	28	10.590	0.9337	60	20
28	29	0.8042	0.7006	120	70
29	30	0.5075	0.2585	200	600
30	31	0.9744	0.9630	150	70
31	32	0.3105	0.3619	210	100
32	33	0.3410	0.5302	60	40

6 Numerical Results

For this computational implementation, the Julia programming environment version 2024a was employed. The computations were conducted on a PC equipped with an AMD Ryzen 7 3700 2.3 GHz processor and 16.0 GB RAM running on

Table 2. Objective function parameters

Par.	Value	Unit	Par.	Value	Unit
C	0.1390	USD/kWh	T	365	Days
Δ_h	0.50	hour	k_1	6/2190	1/Days
k_2	10	years	ω_1	1.50	USD/Mvar^3
ω_2	−713.00	USD/Mvar^2	ω_3	153.750	USD/Mvar

a 64-bit version of Microsoft Windows 10 Single Language. Our proposal was implemented in the Julia software, using the `JuMP` optimization environment and the interior-point optimizer `IPOPT` [10].

Two approaches were compared against it to demonstrate the effectiveness of the SCA-IPOPT optimization methodology. The first involved solving the exact optimization problem via the General Algebraic Modeling System (GAMS) and the BONMIN solver, which is designed to handle MINLP problems. The second approach included several well-known optimization techniques, including the Chu & Beasley Genetic Algorithm (CBGA), PSO, black widow optimization (BWO), and the SCA alone [13].

Remark 3. *In the numerical simulations, two approaches, referred to as SCA-IPOPT$_1$ and SCA-IPOPT$_2$, were considered. The first approach corresponds to the optimal location and sizing of the TSCs with a fixed active power dispatch throughout the operation period, implying a constant reactive power compensation, and the second approach involves varying reactive power injection at each time period.*

Table 3 compares the performance of all the analyzed optimizers.

Table 3. Performance analysis of our proposal and the literature reports used for comparison

Method	Location	Size (Mvar)	f (USD/Year)	Red. (%)
Ben. case	–	–	112740.90	–
BONMIN	[6, 18, 30]	[0.0000, 0.1138, 0.4551]	100221.38	11.10
CBGA	[13, 30, 31]	[0.1528, 0.3227, 0.1157]	100139.21	11.18
PSO	[14, 30, 31]	[0.1486, 0.3244, 0.1157]	100107.24	11.21
BWO	[14, 30, 32]	[0.1486, 0.3337, 0.1064]	100093.29	11.22
SCA	[14, 30, 32]	[0.1486, 0.3337, 0.1064]	100093.29	11.22
SCA-IPOPT$_1$	[14, 30, 32]	[0.1486, 0.3337, 0.1064]	100093.29	11.22
SCA-IPOPT$_2$	[14, 30, 32]	[0.1786, 0.4022, 0.1365]	98729.21	12.43

Table 3 compares the different methods analyzed with regard to the optimal siting and sizing of TSCs in a 33-bus grid, focusing on the last two approaches, *i.e.*, SCA-IPOPT$_1$ and SCA-IPOPT$_2$.

In SCA-IPOPT$_1$, the TSCs are placed at nodes 14, 30, and 32, with sizes of 0.1486 Mvar, 0.3337 Mvar, and 0.1064 Mvar, leading to an annual cost of 100,093.29 USD/year. This represents an 11.22% cost reduction in comparison with the base case. These results are identical to those obtained with the standalone SCA and the BWO [13,20]. While this approach performs well in terms of both cost reduction and placement efficiency, its key advantage lies in the robustness and reliability of the hybrid SCA-IPOPT framework, which offers strong convergence capabilities in solving nonlinear optimization problems.

SCA-IPOPT$_2$, on the other hand, allows for a flexible reactive power injection across time periods, resulting in more refined TSC sizes (0.1786 Mvar, 0.4022 Mvar, and 0.1365 Mvar) at the same nodes. This flexibility leads to a lower annual cost (98,729.21 USD/year) that represents a 12.43% cost reduction, the best performance among all analyzed methods. The ability to adjust the reactive power injection makes SCA-IPOPT$_2$ more effective at reducing costs and improving system performance. In summary, SCA-IPOPT$_2$ outperforms all other approaches, including SCA-IPOPT$_1$, BONMIN, CBGA, PSO, and BWO, as it yields superior optimization results, particularly in terms of cost efficiency and network flexibility.

7 Conclusions and Future Work

This research addressed the optimal siting and sizing of TSCs in electrical distributions through a hybrid approach that combines the sine-cosine algorithm with the interior-point optimizer within a hybrid optimization framework and a discrete-continuous encoding. The integration of the IPOPT significantly improved the performance of the methodology, achieving reductions of up to 12.43% in the annual operating costs compared to the 11.22% achieved by previous approaches. Numerical validations conducted on the classical 33-bus system demonstrated that the proposed hybrid SCA-IPOPT approach outperforms traditional optimizers such as the CBGA and PSO. The results were further validated in comparison with those obtained via the GAMS software and the BONMIN solver - while effective, these results tended to converge to local optima. Additionally, the numerical results obtained using the SCA were comparable to those achieved by the BWO, thus confirming the robustness of both algorithms in solving the complex optimization problem. The findings reaffirm that injecting reactive power through TSCs in electrical distribution networks remains a highly effective strategy for reducing energy losses and enhancing the quality of the electricity service.

Future research could focus on three key areas: (i) accounting for uncertainties in the demand behavior, as well as considering the potential integration of renewable energy resources into electrical distribution grids; (ii) developing a convex approximation of the exact MINLP approach to evaluate its effectiveness compared to the proposed SCA-IPOPT method; and (iii) extending the proposed formulation to power transmission networks while incorporating additional reactive power compensation systems.

Acknowledgments. The authors acknowledge the support provided by the Thematic Network 723RT0150, *Red para la integración a gran escala de energías renovables en sistemas eléctricos* (RIBIERSE-CYTED), funded by the 2022 call for Thematic Networks of the CYTED (Ibero-American Program of Science and Technology for Development). This research was also supported by project no. 6-24-9, titled *Desarrollo de una metodología de control secundario para microrredes de corriente continua aisladas empleando control predictivo basado en el modelo* and attached to Universidad Tecnológica de Pereira's Research, Innovation, and Extension Vice-Principalship.

Disclosure of Interests. The authors declare no competing interests that are relevant to the content of this article.

References

1. Bhattacharyya, B., Basak, S., Rajbhar, S.K.: Sine cosine-based optimization method for reactive power planning with multitype FACTS devices (2024)
2. Biegler, L., Zavala, V.: Large-scale nonlinear programming using IPOPT: an integrating framework for enterprise-wide dynamic optimization. Comput. Chem. Eng. **33**(3), 575–582 (2009). https://doi.org/10.1016/j.compchemeng.2008.08.006
3. Capitanescu, F., Glavic, M., Ernst, D., Wehenkel, L.: Interior-point based algorithms for the solution of optimal power flow problems. Electric Power Syst. Res. **77**(5–6), 508–517 (2007). https://doi.org/10.1016/j.epsr.2006.05.003
4. Chakraborty, S., Mukhopadhyay, S., Biswas, S.: Coordination of D-STATCOM &âĂŕ; SVC for dynamic VAR compensation and voltage stabilization of an AC grid interconnected to a DC microgrid. IEEE Trans. Ind. Appl. **58**(1), 634–644 (2022). https://doi.org/10.1109/tia.2021.3123264
5. Gabis, A.B., Meraihi, Y., Mirjalili, S., Ramdane-Cherif, A.: A comprehensive survey of sine cosine algorithm: variants and applications. Artif. Intell. Rev. **54**(7), 5469–5540 (2021). https://doi.org/10.1007/s10462-021-10026-y
6. Garrido-Arévalo, V.M., Gil-González, W., Montoya, O.D., Grisales-Noreña, L.F., Hernández, J.C.: Optimal dispatch of DERs and battery-based ESS in distribution grids while considering reactive power capabilities and uncertainties: A second-order cone programming formulation. IEEE Access **12**, 48497–48510 (2024). https://doi.org/10.1109/access.2024.3382940
7. Gupta, A.R., Kumar, A.: Energy savings using D-STATCOM placement in radial distribution system. Procedia Comput. Sci. **70**, 558–564 (2015)
8. Jawad, I.M., Abdulrasool, A.Q., Mohammed, A.Q., Majeed, W.S., Salim ALRikabi, H.T.: The optimal allocation of thyristor-controlled series compensators for enhancement HVAC transmission lines Iraqi super grid by using seeker optimization algorithm. Open Eng. **14**(1), 20220499 (2024)
9. Kabalci, E.: Reactive power compensation in AC power systems. In: Mahdavi Tabatabaei, N., Jafari Aghbolaghi, A., Bizon, N., Blaabjerg, F. (eds.) Reactive Power Control in AC Power Systems. PS, pp. 275–315. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51118-4_8
10. Lubin, M., Dowson, O., Dias Garcia, J., Huchette, J., Legat, B., Vielma, J.P.: JuMP 1.0: recent improvements to a modeling language for mathematical optimization. Math. Program. Comput. (2023). https://doi.org/10.1007/s12532-023-00239-3

11. Maataoui, Y., Chekenbah, H., Boutfarjoute, O., Puig, V., Lasri, R.: A coordinated voltage regulation algorithm of a power distribution grid with multiple photovoltaic distributed generators based on active power curtailment and on-line tap changer. Energies **16**(14), 5279 (2023). https://doi.org/10.3390/en16145279
12. Mirjalili, S.: SCA: a sine cosine algorithm for solving optimization problems. Knowl. Based Syst. **96**, 120–133 (2016). https://doi.org/10.1016/j.knosys.2015.12.022
13. Montoya, O.D., Garrido-Arévalo, V.M., Gil-González, W.: Optimal sizing and location of TSCS in medium-voltage distribution networks via the sine-cosine algorithm. In: 2024 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI), pp. 1–6. IEEE (2024). https://doi.org/10.1109/colcaci63187.2024.10666524
14. Montoya, O.D., Gil González, W.J., Grisales-Noreña, L.F., Giral, D., Molina-Cabrera, A.: On the optimal reconfiguration of radial AC distribution networks using an MINLP formulation: a gams-based approach. Ingeniería e Investigación **42**(2), e91192 (2021). https://doi.org/10.15446/ing.investig.91192
15. Montoya, O.D., Ramírez-Vanegas, C.A., González-Granada, J.R.: Dynamic active and reactive power compensation in distribution networks using PV-STATCOMS: a tutorial using the Julia software. Results Eng. **21**, 101876 (2024). https://doi.org/10.1016/j.rineng.2024.101876
16. Nemirovski, A.S., Todd, M.J.: Interior-point methods for optimization. Acta Numer. **17**, 191–234 (2008). https://doi.org/10.1017/s0962492906370018
17. Pal, K., Verma, K., Gandotra, R.: Optimal location of FACTS devices with EVCS in power system network using PSO. e-Prime-Adv. Electr. Eng. Electron. Energy **7**, 100482 (2024)
18. Rauch, J., Brückl, O.: Achieving optimal reactive power compensation in distribution grids by using industrial compensation systems. Electricity **4**(1), 78–95 (2023). https://doi.org/10.3390/electricity4010006
19. Rezaeian Marjani, S., Talavat, V., Galvani, S.: Optimal allocation of D-STATCOM and reconfiguration in radial distribution network using MOPSO algorithm in TOPSIS framework. Int. Trans. Electr. Energy Syst. **29**(2), e2723 (2019)
20. Santamaria-Henao, N., Montoya, O.D., Trujillo-Rodríguez, C.L.: Optimal siting and sizing of FACTS in distribution networks using the black widow algorithm. Algorithms **16**(5), 225 (2023). https://doi.org/10.3390/a16050225
21. Sirjani, R., Jordehi, A.R.: Optimal placement and sizing of distribution static compensator (D-STATCOM) in electric distribution networks: a review. Renew. Sustain. Energy Rev. **77**, 688–694 (2017)
22. Tolabi, H.B., Ali, M.H., Rizwan, M.: Simultaneous reconfiguration, optimal placement of DSTATCOM, and photovoltaic array in a distribution system based on fuzzy-ACO approach. IEEE Trans. Sustain. Energy **6**(1), 210–218 (2014)
23. Vannelli, A., Quintana, V.H., Vargas, L.: Interior point optimization methods: theory, implementations and engineering applications. Can. J. Electr. Comput. Eng. **17**(2), 84–94 (1992). https://doi.org/10.1109/cjece.1992.6592637
24. Wang, Y., Jiang, Q.: Reactive power optimizatin of distribution network based on primal-dual interior point method and simplified branch and bound method. In: 2014 IEEE PES T&;D Conference and Exposition. IEEE (2014). https://doi.org/10.1109/tdc.2014.6863552

Recognizing Overarching Themes and Actors in Peacebuilding: A Longitudinal Analysis of Press Content in Latin America

Aglaya Batz[1(✉)], Merlin P. Grueso-Hinestroza[1], Lizeth F. Serrano-Cardenas[1], David F. D'croz-Baron[2], and Carlos A. Ojeda-Sanchez[1]

[1] Innovation Center - School of Management, Universidad del Rosario, Calle 200 Between Carrera 7 and Autopista Norte, Bogotá D.C., Colombia
aglaya.batz@urosario.edu.co

[2] Artificial Intelligence Research, D&N Business Intelligence and Consulting S.A.S., Bogotá D.C., Colombia

Abstract. Despite extensive efforts by academia and affiliated organizations to broaden the conceptual understanding of peacebuilding, practical approaches in various contexts continue to diverge. These divergences are often influenced by cultural, temporal, and geographic factors. This paper seeks to address the following research question: What cross-cutting themes can be identified in peacebuilding across diverse contexts? To investigate this, a comprehensive statistical topic modeling approach was applied to news articles published in leading newspapers from Colombia, Mexico, Chile, and Peru between 2010 and 2020. The results identified six key thematic areas within the peacebuilding discourse: (1) the regional significance of the peace processes in Colombia; (2) the Catholic Church's role in promoting peacebuilding; (3) the influence of ongoing global wars and conflicts; (4) emphasis on domestic conflicts, both past and present; (5) the private sector's involvement in peacebuilding; and (6) initiatives related to the Culture of Peace and peace education.

Keywords: Peacebuilding · Unsupervised Learning · Content analysis · Computational Social Sciences · Latent Dirichlet Allocation

1 Introduction

Peacebuilding, as defined by Boutros-Ghali [1], involves actions aimed at identifying and strengthening structures that promote lasting peace, with the goal of preventing a relapse into conflict. This endeavor entails developing initiatives that sustain long-term commitments from key actors and gaining insights into the factors that contribute to conflict. Despite its formulation three decades ago, gaps remain in both the operationalization and comprehension of the concept. Boutros-Ghali's definition has often been employed more rhetorically than practically within political discourse [2], underscoring the need for a nuanced understanding of the underlying issues to facilitate its practical application.

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 89–107, 2025.
https://doi.org/10.1007/978-3-031-88854-0_7

Empirical studies have demonstrated that peacebuilding activities and their underlying issues vary across conflict zones, influenced by cultural contexts and the characteristics of the conflicts themselves [3, 4]. Considering this, the present study seeks to address the following question: What cross-cutting themes can be identified in peacebuilding across diverse contexts? This inquiry is situated within two key recognitions: (1) the media plays a pivotal role in constructing and disseminating narratives on conflict, war, and peace to the public [5] and (2) Latin America offers a unique research landscape, characterized by diverse forms of violence and varying conflict intensities.

To explore this question, content analysis was conducted on news articles from major newspapers in Colombia, Mexico, Chile, and Peru, published between 2010 and 2020. Prior research [6] has demonstrated that content analysis can provide valuable insights into peace and peacebuilding beyond regional boundaries. Previous studies have employed news sources to examine peacebuilding challenges using content analysis methodologies, either by focusing on single newspapers within a country [3] or by comparing how topics related to conflict and peace are portrayed across media outlets from countries experiencing tension [4].

Research has highlighted the effectiveness of content analysis, especially when combined with unsupervised learning techniques such as Latent Dirichlet Allocation (LDA), in structuring texts based on latent topics that emerge from word co-occurrences [7–10]. Compared to other content analysis methodologies—such as manual coding, dictionary-based codification, or supervised learning algorithms—LDA offers significant advantages by allowing for a data-driven analysis of large text corpora, facilitating swift information processing and enhanced comprehension of knowledge domains.

A comprehensive comparison of content analysis methodologies, originally provided by Quinn, K.M. et al. [8], is summarized in Table 1. It demonstrates that topic modeling methods, such as LDA, provide considerable benefits when analyzing large collections of texts. However, scholars have also noted limitations to these methods, emphasizing the importance of the researcher's deep understanding of the subject to ensure that results are properly contextualized. Furthermore, they advocate for rigorous documentation procedures throughout the research process, including data extraction, exclusion, refinement, and analysis [11].

The subsequent sections of this paper outline each methodological step, present country-specific breakdowns of the results, discuss findings in relation to the identified concerns, and conclude with insights derived from the study.

2 Methodology

A methodological approach grounded in standard content analysis practices, incorporating semantic algorithms [11], was employed to conduct a thorough semantic and longitudinal analysis of news articles related to peacebuilding from major newspapers in Colombia, Mexico, Chile, and Peru.. This methodology is outlined in Fig. 1 and as follows:

Table 1. Comparison of Content Analysis methodologies employed in text analysis. Source: [8]

	Method				
	Human reading	*Human Coding*	*Dictionaries*	*Supervised Learning*	*Topic Models*
Assumptions					
Categories are known	NO	YES	YES	YES	NO
Category nesting, if any, is known	NO	YES	YES	YES	NO
Relevant text features are known	NO	NO	YES	YES	YES
Mapping is known	NO	NO	YES	NO	NO
Coding can be automated	NO	NO	YES	YES	YES
Costs					
Pre-analysis Costs					
• Person-hours spent conceptualizing	Low	High	High	High	Low
• Level of substantive knowledge	Moderate/high	High	High	High	Low
Analysis Costs					
• Person-hours spent per text	High	High	Low	Low	Low
• Level of substantive knowledge	Moderate /high	Moderate	Low	Low	Low
Post-analysis Costs					
• Person-hours spent interpreting	High	Low	Low	Low	Moderate
• Level of substantive knowledge	High	High	High	High	High

2.1 Retrieving Relevant News

For each country, a prominent newspaper offering a digital archive through academic databases, or its website was selected. Specifically, the **vLex Global** database was used to download the headlines and full texts of relevant news articles in PDF format. Articles related to peacebuilding published in **El Tiempo** (Colombia), **El Universal** (Mexico), **El Mercurio** (Chile), and **El Comercio** (Peru) between October 1, 2010, and December 31, 2020, were retrieved from vLex Global using search queries (1) to (4). All search queries were conducted in Spanish, the original language of the newspapers:

$$\text{constru} * \text{ AND paz} \tag{1}$$

$$\begin{gathered}[\text{construccion paz}]/10 \text{ OR } [\text{construir paz}]/10 \text{ OR } [\text{construirá paz}]/10 \text{ OR } [\text{construya paz}]/\\ 10 \text{ OR } [\text{construye paz}]/10 \text{ OR } [\text{construyen paz}]/10 \text{ OR } [\text{construyendo paz}]/10\end{gathered} \tag{2}$$

$$\begin{gathered}[\text{construcción paz}]/10 \text{ OR } [\text{construir paz}]/10 \text{ OR } [\text{construya paz}]/\\ 10 \text{ OR } [\text{construye paz}]/10 \text{ OR } [\text{construyendo paz}]/10\end{gathered} \tag{3}$$

$$\begin{gathered}[\text{consolidar paz}]/10 \text{ OR } [\text{consolidación paz}]/10 \text{ OR } [\text{consolide paz}]/10 \text{ OR } [\text{consolida paz}]/\\ 10 \text{ OR } [\text{consolido paz}]/10 \text{ OR } [\text{consolidaron paz}]/10 \text{ OR } [\text{consolidando paz}]/10\end{gathered} \tag{4}$$

Initially, every news article containing *construcción*, *construir*, *construye(n)[do]* and *paz* in their headlines or texts—regardless of word order or distance—were matched through query (1). However, articles where both terms were separated by more than ten words or pertained to subjects unrelated to peacebuilding were manually excluded. This initial prefilter reduced the dataset to 1,943 relevant articles in Colombia and 724 in Mexico.

However, for Chile and Peru, search strategy (1) happened to be too broad, resulting in numerous irrelevant matches. Hence, further queries were defined to search in these countries, where search terms were separated between them by no more than ten words, with the drawback of employing all conjugations of verbs *construcción* and *consolidación* therein, because of vLex Global limitations. Additionally, the Spanish sentence *Consolidación de paz*, itself another translation of peacebuilding to Spanish, was considered for the latter equations. Overall, search strategies following these criteria, that is, (2) to (4), matched 130 relevant news articles in Chile and 80 in Perú.

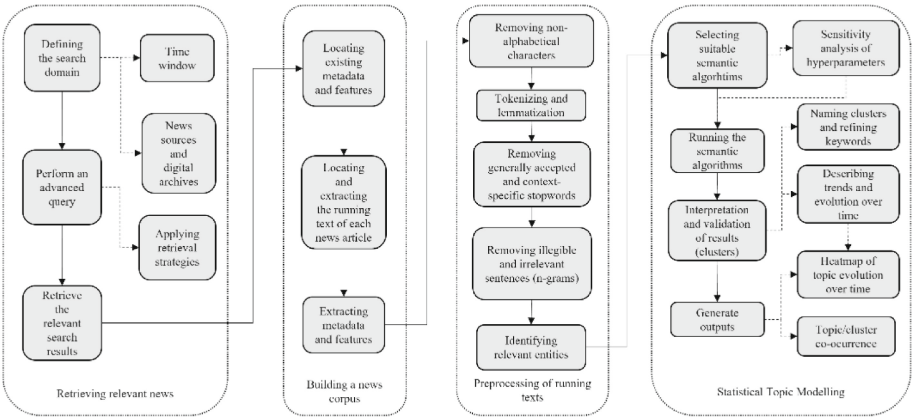


Fig. 1. Methodological approach to content analysis of (peacebuilding-related) news articles

After extracting the articles' texts, 300 articles from *El Tiempo* were excluded due to their authors being editorialists or columnists, as this study focuses on news coverage rather than opinion pieces. No such exclusions were necessary for the other newspapers.

2.2 Building a News Corpus

The running text and metadata from the downloaded news articles were extracted using the *PDFminer.six* and *Selenium* libraries for Python. Each country's dataset was treated as an independent corpus, and four Comma-Separated Value (CSV) databases were constructed, containing the headlines, publication dates, and running texts of the news articles.

2.3 Preprocessing of Running Texts

Before conducting statistical topic modeling, the news texts were cleaned of irrelevant words, noise, grammatical connectors, general and context-specific stopwords, and Spanish-language expressions related to peacebuilding. For this purpose, the software libraries *clean-text*, *spaCy*, and *NLTK* [12], available for the Python Programming Language were used.

clean-text removed nonalphabetic symbols from the running texts, such as punctuation signs, numbers, and digits, as well as e-mail and website addresses. *SpaCy* was used to tokenize and lemmatize the running texts as follows: Tokenizing decomposes or splits a text into words or sentences. In this work, texts were tokenized at the word level. By lemmatizing, all conjugations of Spanish-language verbs are replaced by their corresponding root. *SpaCy* was of special interest in this work, slightly deviating from already-established, well-proven Natural Language Processing (NLP) procedures and software libraries, due to its ability to lemmatize Spanish verbs with outstanding accuracy. Lastly, *NLTK* was used for obtaining the most relevant n-grams from running texts. An n-gram is a group of two to n words -in this case, up to four words-, and usually reveals unfiltered noise, or relevant names of people, organizations, companies, places, legislation, and similar -entities-.

N-grams containing two, three, or four words were analyzed and cleaned using a two-pass procedure. The first pass typically revealed additional noise and occasionally identified relevant sentences that exceeded four words in length. This initial step was instrumental in reducing ambiguity within the data. The second pass of n-gram cleaning then refined the results, ensuring that any remaining noise was eliminated while preserving essential content. Expert judgment and prior knowledge were critical during this process, particularly in relation to peacebuilding, conflicts, wars, Latin American history, and the political, social, and economic contexts of each country analyzed. This expertise was essential to making informed cleaning decisions and ensuring that important sentences were accurately identified and retained for further analysis.

2.4 Statistical Topic Modeling

The *Latent Dirichlet Allocation* (LDA) is a probabilistic and generative model for collections of discrete data, such as the running text of news items, widely used to uncover their hidden semantic content [13]. In this model, the terms contained in analyzed texts come from a mixture of non-correlated groupings distributed over the vocabulary of the corpus, and it is generative since it assigns the words of a document to random variables with Dirichlet distribution, performing the semantic grouping through a recurrent probabilistic process [13]. LDA is recurrently used in social sciences and related disciplines due to its scalability, reliability, and efficiency [7, 11, 14, 15].

The *gensim* library [16] for the Python Programming Language, provided the implementation of the LDA algorithm used in this work. LDA hyperparameters, such as the number of topics -clusters-, passes, and iterations, are arbitrary in nature and independent of features and attributes in the corpora. Hence, a grid-search sensitivity analysis of the hyperparameters was performed, where the resulting LDA model with the highest coherence was selected as the definitive model. The coherence metric used in this work was based upon the work detailed in [17].

As mathematical models often struggle to capture the nuanced meaning of text, expert validation was required to assess the quality of the initial LDA clusters [18]. Nonetheless, to improve intelligibility, LDA clusters can be further refined using the term relevance model detailed in [18] and Eq. (5), where the arbitrary λ parameter ($0 \leq \lambda \leq 1$) is "the weight given to the probability of term w under topic k (ϕ_{kw}) relative to its *lift* (measuring both on the log scale)", and *lift* is the (log-scaled) ratio of probabilities of w across k -ϕ_{kw}- and the corpus (p_w).

$$r(w, k|\lambda) = \lambda \log \phi_{kw} + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right) \tag{5}$$

We opted for refining the initial LDA clusters using this model, implemented in the *pyLDAvis* library for the Python Programming Language, and following the recommendations stated in [18], λ was set to 0.61. For each country studied, refined LDA clusters were validated by experts from the top 30 most relevant keywords and news share distribution over time for each cluster and cross-checked with random samples of news articles from the corpus.

3 Results

3.1 Colombia (El Tiempo)

In Colombia, the analysis focused on 1,625 peacebuilding-related news articles obtained from **El Tiempo**. These articles were examined to describe their temporal evolution, as illustrated in Fig. 2. The time frame analyzed corresponds with the peace process known as the "*Acuerdo Final para la Terminación del Conflicto y la Construcción de una Paz Estable y Duradera*", a peace agreement reached between the Colombian government and the Revolutionary Armed Forces of Colombia (FARC) after four years of negotiations starting in 2012. Additionally, the latest peace talks between the Colombian government and the National Liberation Army (ELN) fall within this period.

The peak in peacebuilding-related news coverage corresponds to the October 2016 referendum, which failed to ratify the peace agreement between the Colombian government and FARC. Following this, coverage and interest in peacebuilding gradually diminished, reflecting decreased concern over potential breaches of the peace agreement by the FARC. This decline suggests that influential interest groups, such as businesses, tend to disengage from peace-related issues unless disruptions caused by FARC's activities re-emerge.

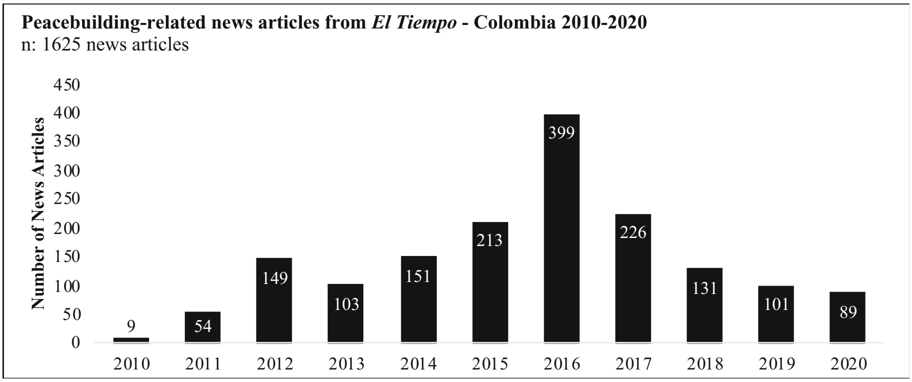


Fig. 2. Temporary evolution of peacebuilding-related news coverage from El Tiempo, per year

The topic model for Colombia identified **eleven clusters**, based on the most relevant keywords in the peacebuilding-related articles. A sample of the first three clusters is displayed in Table 2, covering themes such as victims' grievances, reconciliation, post-conflict and rural development, peace talks with FARC and ELN, transitional and restorative justice, international cooperation, disarmament, demobilization, reintegration (DDR), historical memory, political compromises, and the development of a Culture of Peace.

Within the eleven clusters, four main themes emerge:

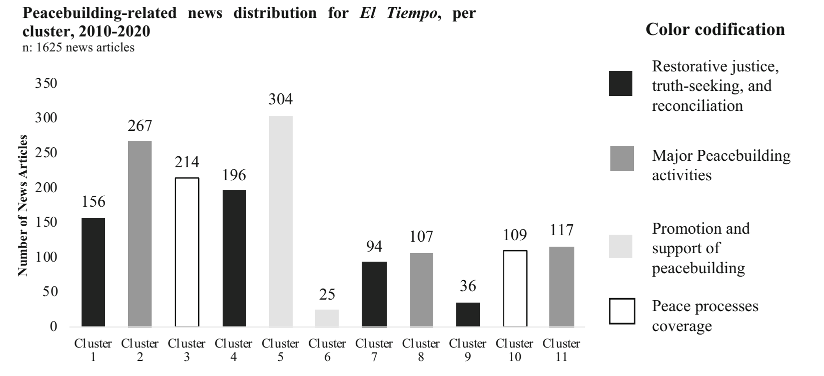
1. **Restorative Justice and Reconciliation**: Represented by Clusters 1, 4, 7, and 9, this theme includes truth-seeking, trust-building, reconciliation, and the roles of institutions such as the Special Jurisdiction for Peace (JEP) and the National Center for Historical Memory (CNMH).
2. **Peacebuilding Activities**: This theme, spanning Clusters 2, 8, and 11, addresses post-conflict development, DDR, and the promotion of a Culture of Peace.
3. **Political Evolution and Peace Negotiations**: Clusters 3 and 10 cover political milestones in peace negotiations with FARC and ELN, highlighting instances where peacebuilding is mentioned without clear definitions or actions.
4. **Professional and Public Perspectives on Peace Process**: Clusters 5 and 6 encompass views from professionals, business leaders, and public figures on peace, peaceful conflict resolution, and peacebuilding initiatives.

Approximately 60% of the news articles from **El Tiempo** focus on victims' needs, DDR, transitional justice, and grassroots peacebuilding efforts related to the Culture of Peace as depicted in Fig. 3.

The heatmap in Fig. 4 further illustrates the temporal evolution of these clusters, indicating shifts in coverage over the analyzed years. The heatmap shows increasing

Table 2. Most relevant keywords describing the first three LDA clusters from peacebuilding-related news published in El Tiempo, 2010–2020

#	Cluster 1	Relevance	Cluster 2	Relevance	Cluster 3	Relevance
1	*Vida* (life)	−2,28017	*Desarrollo* (Development)	−2,32831	ELN	−1,82521
2	*Vivir* (living)	−2,5153	*Proyecto* (Project)	−2,34785	Guerrilla	−1,94939
3	*Hijo* (Child)	−2,52363	*Tierra* (Land)	−2,37401	*Militar* (Military)	−2,13467
4	*Casa* (House)	−2,53413	*Programa* (Program)	−2,40736	*cese_fuego* (Ceasefire)	−2,19102
5	*Pueblo* (town)	−2,55649	*Recurso* (Resource)	−2,44714	*Negociación* (Negotiation)	−2,25503
6	*Padre* (father)	−2,59981	Rural	−2,44849	*Habana* (Havana)	−2,28701
7	*Familia* (Family)	−2,61134	*Empresa* (Enterprise)	−2,468	*Diálogo* (Conversation)	−2,35741
8	Gente (people)	−2,61805	*Posconflicto* (Postconflict)	−2,57698	*Mesa* (Roundtable)	−2,36392
9	*Iglesia* (Church)	−2,65253	Region	−2,58554	*Jefe* (Chief)	−2,45686
10	*Querer* (want, love)	−2,72824	Social	−2,60465	*Guerrillero* (Guerrilla Warrior)	−2,47906
11	*Ir* (Go)	−2,73843	*Campo* (Farmland, Countryside)	−2,60614	*proceso_paz* (Peace Process)	−2,53326
12	*Matar* (Kill, assassinate)	−2,7598	*Inversión* (Investment)	−2,63921	*Ejército* (Army)	−2,54849
13	*Sentir* (Feel)	−2,78293	*Productivo* (Productive)	−2,80038	Bilateral	−2,61201
14	*Papa* (Pope)	−2,80532	*Ciento* (Hundreds)	−2,83321	*Parte* (Party)	−2,65238
15	*Dolor* (Pain, Grief)	−2,84451	*Económico* (Economical)	−2,88332	*fuerza_público* (Law Enforcement)	−2,70619



Cluster 1 – Victims' grievances and callings for reconciliation; Cluster 2 - Postconflict and rural development; Cluster 3 - Coverage of peace talks with FARC and ELN; Cluster 4 - Implementing transitional and restorative justice in Colombia; Cluster 5 – Promotion and support of peacebuilding and peaceful conflict resolution; Cluster 6 – International cooperation for peacebuilding and postconflict activities; Cluster 7 - Milestones of transitional and restorative justice in Colombia; Cluster 8 - DDR and narcotraffic influence on Colombian armed conflict; Cluster 9 – Constructing the historical memory of Colombian armed conflict; Cluster 10 - Negotiating a political compromise for peace; Cluster 11 - *Cultura de Paz* in education and by grassroots movements

Fig. 3. Cluster distribution of Peacebuilding-related news coverage from El Tiempo

coverage in **Cluster 7** as transitional justice milestones are achieved, notably following the final report of the Truth-Seeking Commission. Interest in **Cluster 8** peaked during the DDR process for FARC ex-combatants in 2017 but has declined since. **Cluster 2**, which covers post-conflict and rural development, received significant coverage during peace talks and in response to government policies, with occasional resurgence in recent years, particularly in the context of the COVID-19 pandemic and subsequent analyses of the peace process.

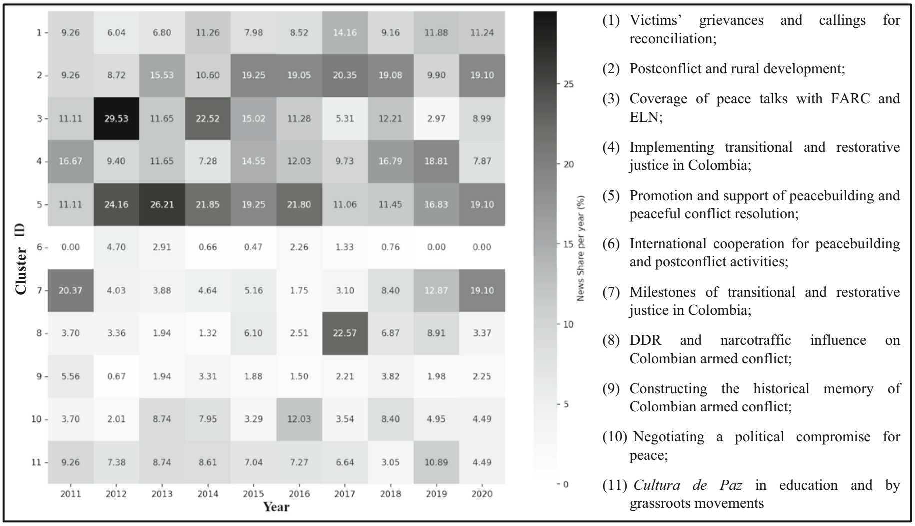


Fig. 4. Temporary evolution of peacebuilding-related news coverage from El Tiempo, by clusters, 2011–2020

3.2 Mexico (El Universal)

Mexico's peacebuilding-related news coverage from *El Universal* comprises a corpus of 701 articles, the temporal evolution of which is depicted in Fig. 5. Given Mexico's ongoing drug war and efforts to resolve it peacefully, the country holds particular significance for this study. Its experience with conflict may shed light on how peacebuilding is framed in other conflict-affected nations.

The overall trend in Mexican news articles shows an upward trajectory, with a notable increase in 2011 following the tragic murder of Javier Sicilia's son, a poet and journalist in Mexico. Sicilia subsequently founded the grassroots organization *Movimiento por la Paz con Justicia y Dignidad* (Movement for Peace with Justice and Dignity), advocating for large-scale demonstrations against injustice and brutality in Mexico's drug war. The movement frequently invoked concepts related to peacebuilding such as Safety Sector Reform (SSR), victim compensation, and respect for human rights.

The late increase in news coverage correlates with the establishment of Local Committees for Peacebuilding and Security in each Mexican state under President Andres

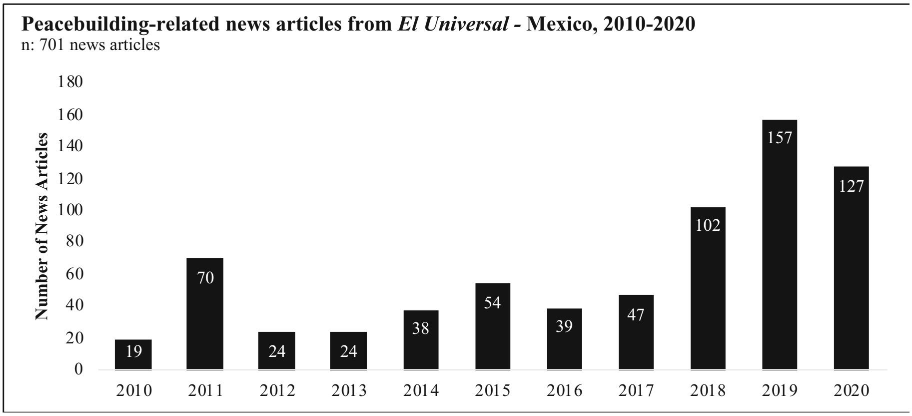


Fig. 5. Temporary evolution of peacebuilding-related news coverage from El Universal, per year

Manuel Lopez Obrador's administration. These committees aimed to facilitate dialogue between the public and state and local authorities in a peaceful, orderly, and democratic environment. However, their focus has shifted from peacebuilding efforts to discussions on public safety and security issues like kidnappings, killings, and gang activities.

Applying Latent Dirichlet Allocation (LDA) to the *El Universal* corpus identified nine clusters, where Clusters 1 and 9 underscore the Catholic Church's role in advocating for peaceful conflict resolution, including setting up peace talks with drug gangs and acting as mediators to address social grievances. Cluster 5 encompasses political and legislative debates on Security Sector Reform (SSR), victim compensation, and other peacebuilding-related topics. Clusters 4 and 8 address criminal and gang activities denounced by the peacebuilding committees, while Cluster 3 contains news and press releases related to these committees. Social issues such as gender-based violence and support for victims are covered in Clusters 2 and 7, and Cluster 6 focuses on promoting the Culture of Peace in classrooms and through grassroots movements.

Figure 6 illustrates the distribution of peacebuilding-related news among these clusters, highlighting political and legislative debates (Cluster 5) and news related to the Peacebuilding committees (Clusters 3 and 8) as dominant themes. Cluster 1 and 9, centered on the role of the Catholic Church, were dominant in 71 news articles, while issues like gender-based violence (Cluster 7) and victim-related matters (Cluster 2) were discussed in 61 news articles. The Peacebuilding committees' news coverage, concentrated in later years, comprises 138 articles (Cluster 3) and 219 articles when considering Cluster 8's relevance to this theme.

The temporal evolution of the nine news clusters for Mexico is detailed in Fig. 7, showing that political and legislative debates on peacebuilding initiatives (Cluster 5) were the dominant theme in news coverage. Cluster 6, promoting the Culture of Peace, peaked in 2016 and 2017 unrelated to specific historical events, while the Peacebuilding committees' coverage increased notably in later years.

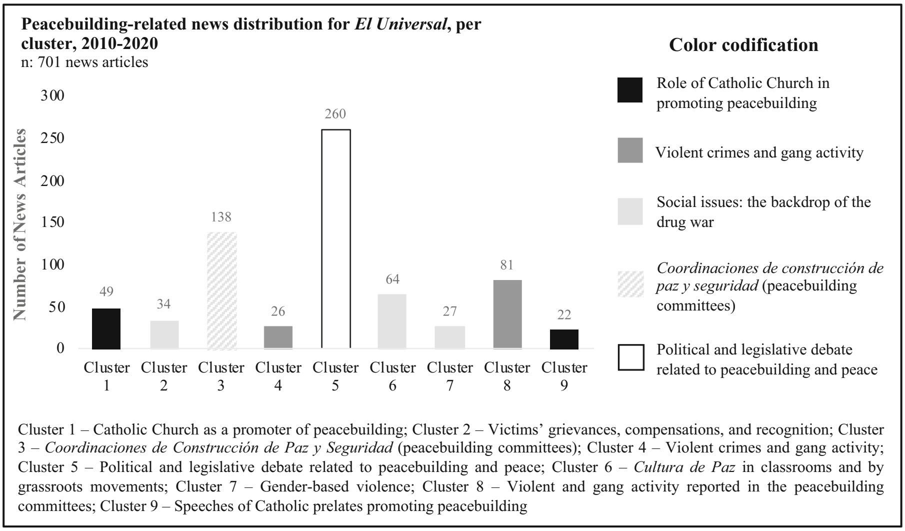


Fig. 6. Cluster distribution of Peacebuilding-related news coverage from El Universal

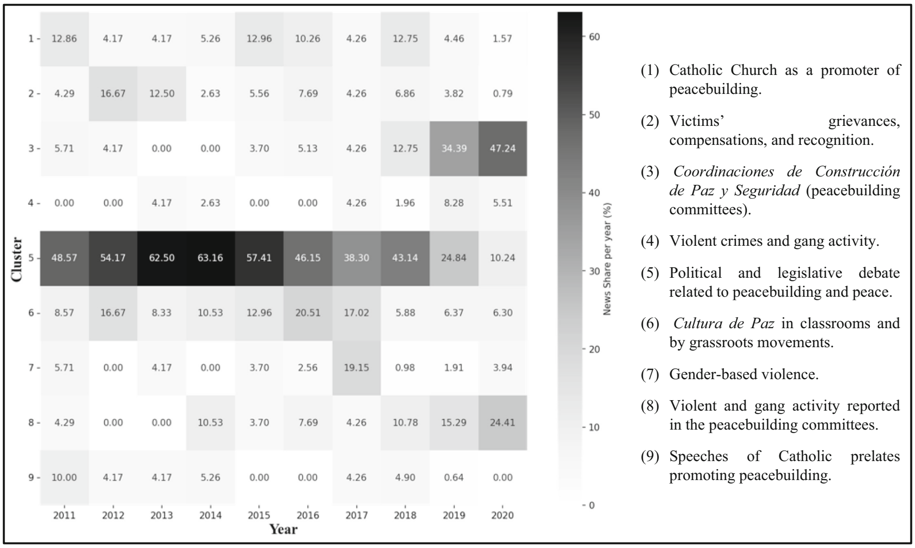


Fig. 7. Temporary evolution of peacebuilding-related news coverage from El Universal, by clusters, 2011–2020

3.3 Chile and Peru

For Chile and Peru, 127 articles from El Mercurio and 77 from El Comercio were analyzed, respectively, describing the temporary evolution as depicted in Fig. 8. Although these nations did not experience wars or military conflicts during the period under consideration, they have a history of disagreements and conflicts with each other and neighboring nations. However, two continuous social disputes and unrest in Chile—the Mapuche conflict since the 19th century and the *Estallido Social*, including the student protests of 2011—are notable. Understanding whether peacebuilding has been suggested, advocated, or attempted to address these issues in Chile is crucial.

According to the volume of news, this has indeed been the case, albeit through a few news articles, as revealed in Fig. 9. Furthermore, a specific cluster (cluster 2) is dedicated to the Mapuche conflict, which has the most peacebuilding-related news coverage in El Mercurio, with 30 news articles. The *Estallido Social* situation might be within the coverage of miscellaneous local news (cluster 5), which includes news articles and interviews covering more than one matter, revealing that peacebuilding in this context is not dominant enough, text-wise, to be concentrated around a dedicated cluster. Instead, the death of Ernesto Videla in 2013, chief negotiator of the Treaty of Peace and Friendship of 1984 between Chile and Argentina, emerges as a small cluster (cluster 7) consisting of five news remembering his legacy and achievements as a peacebuilder.

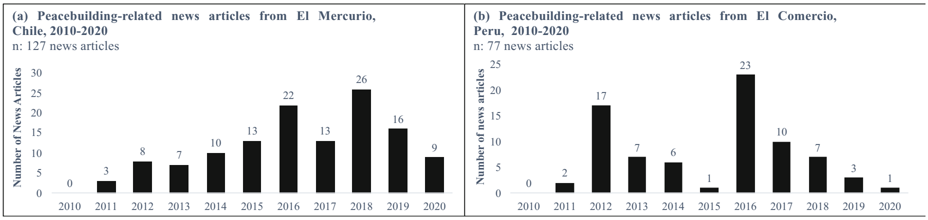


Fig. 8. Temporary evolution of peacebuilding-related news coverage from (a) El Mercurio (Chile) and (b) El Comercio (Peru), per year

Overall, peacebuilding-related news coverage in Chile is mostly international. Themes such as the Colombian peace processes of 2010–2020 (cluster 6), the Arab-Israeli conflict (cluster 4), and Chile's role in peacekeeping and peacebuilding initiatives and operations worldwide (cluster 1) emerge as relevant. These three clusters comprise 69 news articles, as pictured in Fig. 9, detailing the news distribution in the seven clusters revealed. Furthermore, when including a miscellaneous world news cluster containing 6 more articles, the news coverage dedicated to foreign matters increases to 75 articles, about 60% of the news analyzed. The Colombian peace processes might be of particular interest in Chile, as it facilitated peace talks with FARC and ELN.

The temporary evolution of the seven news clusters for Chile is detailed in Fig. 10, in heatmap form. Clusters 4 and 5 are transversal to the time window analyzed and thus are not appropriate to describe or estimate trends for them.

In the case of Peru, the 1980s and 1990s war on terrorism was of particular interest in the semantic analysis of peacebuilding news. Three LDA clusters were found, where Cluster 1 contained most news related to the debate surrounding both Shining Path's (*Sendero Luminoso*) and Alberto Fujimori's government crimes and atrocities, including those against humanity, as well as other domestic issues where peacebuilding is barely mentioned. Most complaints related to the fight against terrorism have already been addressed before the examined period, and the news coverage examined reflects this.

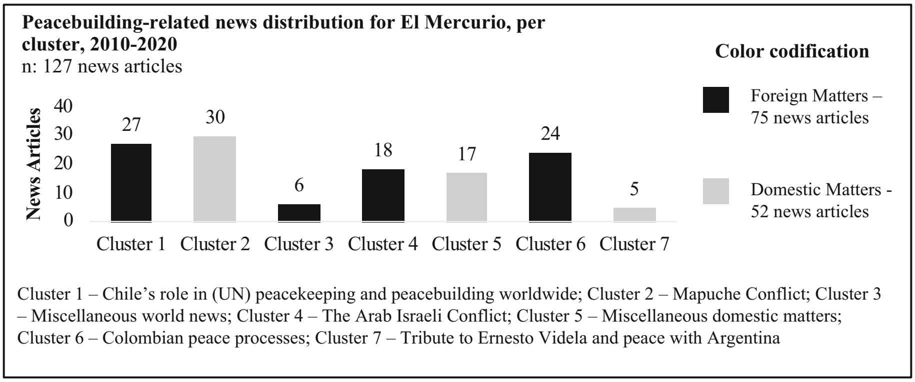


Fig. 9. Cluster distribution of Peacebuilding-related news coverage from El Mercurio

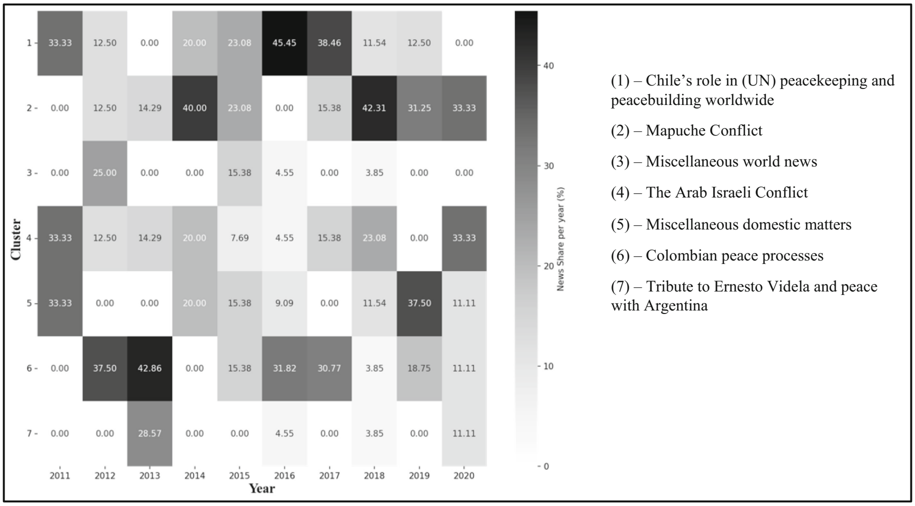


Fig. 10. Temporary evolution of peacebuilding-related news coverage from El Mercurio, by clusters, 2011–2020

Cluster 2 describes a similar domestic theme related to commemorations and resolution of outstanding matters relating to past border conflicts and wars with Ecuador and Chile, such as the Cenepa War of 1995 or the Saltpeter War (*Guerra del Pacífico*) in the 19th century, and Cluster 3 describes the news coverage of the Colombian peace processes with FARC and ELN from the time window analyzed. Figure 11 shows the distribution of total news among the three clusters in detail.

Finally, the temporary evolution of news clusters for Peru is detailed in Fig. 12. The cluster peaks of 2011, 2015, and 2020 are irrelevant, as there is only one peacebuilding-related news article in each of these. Peaks of Cluster 3 relate to key milestones in the 2010s Colombian Peace process. Cluster 2's longitudinal behavior reflects the evolution and dynamics of the maritime border dispute between Peru and Chile that was resolved

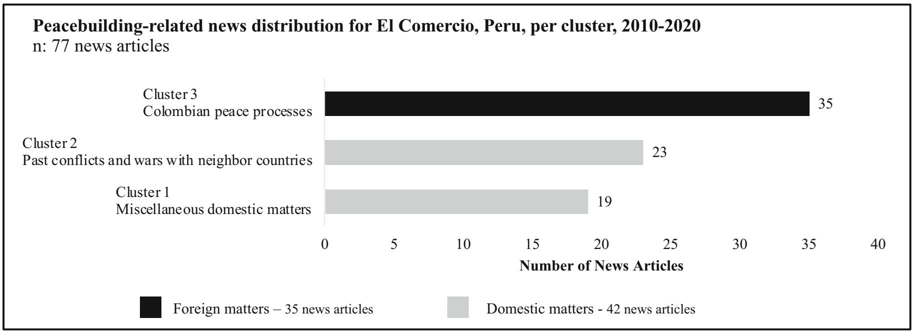


Fig. 11. Cluster distribution of Peacebuilding-related news coverage from El Comercio

by the International Court of Justice in 2014. Otherwise, historical accounts of past conflicts in El Comercio are spontaneous and transversal to this cluster.

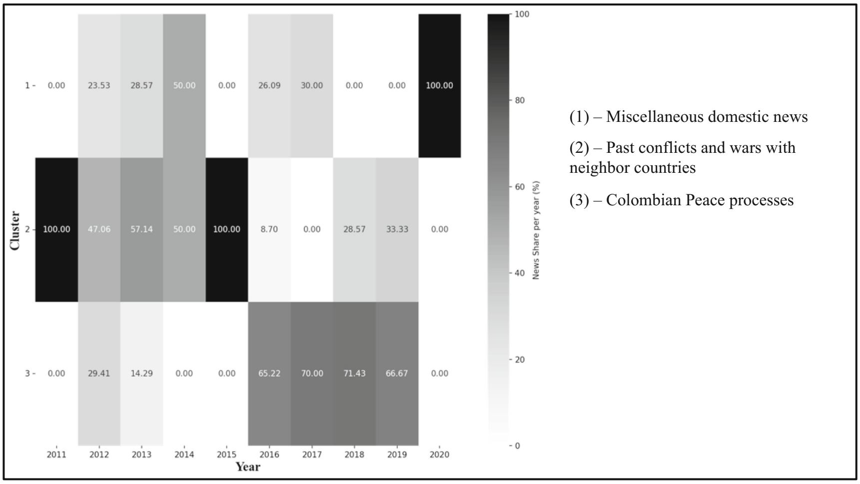


Fig. 12. Temporary evolution of peacebuilding-related news coverage from El Comercio, by clusters, 2011–2020

4 Discussion

The primary objective of this comparative analysis was to identify commonalities and differences in the scope and depth of peacebuilding-related news coverage in four Latin American countries: Colombia, Mexico, Chile, and Peru. This analysis focused on the coverage provided by major, influential newspapers. The findings revealed several key insights: (1) a generally low volume of news coverage on peacebuilding topics, even in the country with the highest coverage, Colombia, which had 1,625 articles over ten years; (2) limited depth in the articles addressing peacebuilding; and (3) notably minimal coverage in Chile and Peru.

Countries currently experiencing ongoing conflicts, such as Mexico and Colombia, displayed a broader and more diverse range of peacebuilding-related themes in their press coverage. In contrast, the press in Chile and Peru tends to focus on fewer peacebuilding-related topics, often relegating them to international events. In both countries, for instance, there are dedicated clusters discussing the Colombian peace process, but these are embedded within a broader assortment of international peace and conflict-related news. Despite two ongoing domestic social conflicts, Chile's newspaper of record, *El Mercurio*, lacks extensive coverage of peacebuilding-related matters. Whether this limitation arises from editorial policies or ideological considerations is beyond the scope of this study.

In the broader context of Latin America, after analyzing the semantic clusters across these four newspapers, several common peacebuilding themes emerged. These commonalities are summarized in Table 3, where themes related to peacebuilding are "shared" when covered, at least, by two newspapers from different countries. Whenever a theme is developed by at least one LDA cluster from a newspaper, the corresponding cell in Table 3 is marked as YES, and where a theme is covered in part by at least one cluster, the cell is marked as Partially. Otherwise, the cell is marked as NO.

Table 3. Summary of shared themes of peacebuilding in the four countries analyzed

News source	Colombian peace processes	Catholic Church as a promoter of peacebuilding	Ongoing wars and conflicts worldwide	Ongoing or former domestic conflicts	The Private Sector's role in peacebuilding	*Culture of Peace* or peace education
El Tiempo (Colombia)	YES (all clusters)	Partially (related to reconciliation – cluster 1)	NO	YES (ongoing, all clusters)	Partially (clusters 2 and 11)	YES (cluster 11)
El Universal (Mexico)	NO	YES (clusters 1 and 9)	NO	YES (ongoing, all clusters)	Partially (cluster 6)	YES (cluster 6)
El Mercurio (Chile)	YES (cluster 6)	NO	YES (clusters 3 and 4)	YES (ongoing, clusters 2 and 5)	NO	NO
El Comercio (Peru)	YES (cluster 3)	NO	NO	YES (former, cluster 1[a])	NO	NO

[a] No news coverage of recent activity by remnants of Shining Path or MRTA, was found among peacebuilding-related articles

Based on the expert interpretation of the Latent Dirichlet Allocation (LDA) clusters, several common themes, depicted in Table 3, were identified:

1. **Colombian Peace Processes:**
   This theme encapsulates the attention given by the Latin American press to the milestones, developments, achievements, and setbacks of the peace talks held in Colombia between 2012 and 2018. Apart from Colombia, where these negotiations were a significant domestic concern, newspapers in Peru and Chile provided substantial coverage of the Colombian peace processes. This coverage stands out against the backdrop of ordinary peacebuilding and foreign affairs reporting, indicating the broader regional significance of Colombia's peace initiatives.
2. **Catholic Church as a Promoter of Peacebuilding:**
   This theme highlights the Catholic Church's involvement in peacebuilding efforts, whether through advocacy in speeches or through more tangible facilitation efforts. For instance, in Mexico's **Clusters 1** and **9**, the Church is seen promoting reconciliation and victim compensation—beyond financial means—related to the Colombian armed conflict. The Church's influence is particularly evident in **Cluster 1**, which addresses these themes directly.
3. **Ongoing Wars and Conflicts Worldwide:**
   Coverage of foreign armed conflicts, wars, or aggressions outside Latin America is also a notable theme. In countries like Mexico and Colombia, internal conflicts often overshadow news on foreign wars. Conversely, in conflict-free countries such as Peru and Chile, this theme features more prominently. In Chile, for example, peacebuilding news in **El Mercurio** often centers on the country's diplomatic traditions, including its involvement in peacekeeping operations and its facilitation of peace talks.
4. **Ongoing or Former Domestic Conflicts:**
   This theme captures news coverage of internal conflicts, both historical and contemporary, regardless of the level of violence. Examples include the Colombian armed conflict, Mexico's drug war, the Mapuche conflict in Chile, and Peru's anti-terrorist campaign during the regime of Alberto Fujimori. Notably, the Colombian peace process of the 2010s attracted more attention from Peruvian and Chilean newspapers than other foreign peacebuilding-related themes. In these contexts, peace initiatives were often framed through the lens of ongoing conflicts, reflecting a normative approach that emphasized the interactions between governments, political opposition, and armed groups [19]. In contrast, news coverage in Peru of past conflicts—resolved either peacefully or through military means over a decade ago—has significantly diminished, suggesting a shift in societal focus toward non-conflict-related issues, typical of more peaceful nations.
5. **The Private Sector's Role in Peacebuilding:**
   This theme addresses the involvement of private enterprises, entrepreneurs, and investors in peacebuilding activities or in promoting peaceful conflict resolution. Although no newspaper clusters were dedicated solely to this theme, private sector contributions to peacebuilding were occasionally mentioned, often as press releases or within regular news coverage. For instance, in Colombia, relevant activities appear in **Clusters 2** and **11**, and in Mexico, similar themes are reflected in **Cluster 6**. However, the role of the private sector in peacebuilding remains marginal and understated in

both countries. Whether this reflects a lack of interest or knowledge on the part of private enterprises warrants further investigation in future research.

6. **Culture of Peace and Peace Education:**

   This theme encompasses peacebuilding activities and the promotion of peaceful conflict resolution by educational institutions (schools, colleges, universities) and grassroots movements. "Peace education" refers to all programs and initiatives aimed at integrating peaceful conflict resolution into curricula, while broader Culture of Peace activities involve civil society efforts to foster peace from the ground up. In conflict-ridden countries like Colombia and Mexico, these initiatives received significant press coverage, forming dedicated clusters that included 117 and 64 news articles, respectively. These articles highlighted peacebuilding efforts led by NGOs, grassroots movements, individuals, communities, and—less frequently—the private sector. In more peaceful societies such as Peru and Chile, however, the Culture of Peace and peace education are less commonly associated with peacebuilding efforts.

5 Conclusions

This study employed content analysis to explore peacebuilding-related news coverage in four Latin American countries—Colombia, Mexico, Chile, and Peru—focusing on prominent media outlets such as *El Tiempo* (Colombia), *El Universal* (Mexico), *El Mercurio* (Chile), and *El Comercio* (Peru). The analysis revealed key insights into the communicative strategies and thematic focus of these newspapers. While this study did not directly analyze communicative strategies, understanding these strategies could significantly enhance the interpretation of the clusters identified in this research. Future studies should employ qualitative approaches, such as discourse analysis or news frame analysis, to provide deeper insights into the biases, frames, and communicative intentions within each newspaper's reporting. This could reveal how the political agendas and editorial policies of media organizations shape the narrative around peacebuilding, influencing public opinion on the matter.

A critical observation from this study was the variation in the portrayal of peacebuilding actors across the media. While some outlets emphasize the role of governments, others focus on corporations, religious organizations, or civil society actors. Understanding the contributions and impacts of these different actors is essential for future research, as it could clarify their roles in fostering peacebuilding ecosystems and highlight collaborative efforts across sectors in addressing social challenges.

The intersection of the private sector and peacebuilding is an area of growing interest in both academic and public spheres. This study found limited coverage of the private sector's role in peacebuilding, suggesting that this relationship warrants further exploration. Future research should examine how the private sector contributes to peacebuilding, the scope of its involvement, and the challenges and limitations it faces. Understanding these dynamics will be critical for formulating comprehensive peacebuilding strategies that involve all societal actors.

Though the political orientations and biases of the analyzed newspapers were beyond the scope of this work, it is important to note that the outlets studied belong to the Pan-American media consortium, *Grupo de Diarios de América* (GDA). This affiliation could influence editorial decisions and coverage priorities. Future research could cross-validate these findings by analyzing news corpora from competitor outlets, such as *El Espectador* in Colombia, *Reforma* in Mexico, *La Tercera* in Chile, and *La República* in Peru, to gain a more balanced understanding of media portrayals of peacebuilding.

The LDA (Latent Dirichlet Allocation) semantic analysis used in this study provided valuable insights but is limited by the model's inability to explicitly account for relationships between clusters or incorporate historical context into the analysis. The interpretation of LDA results is dependent on the expertise of researchers, which was crucial in this study given the social, political, and historical complexities of the Latin American context. Future computational social science projects will require even more sophisticated interdisciplinary collaboration between computer scientists, data scientists, and social scientists. This is particularly true as the use of machine learning (ML) and natural language processing (NLP) algorithms continues to expand, including the emerging trend of using large language models (LLMs) in the analysis of non-numerical data. The challenges of integrating mathematical models with the nuances of text data are significant, underscoring the need for domain expertise to interpret outputs meaningfully.

Further research should explore the application of advanced semantic algorithms, such as **Dynamic Topic Models (DTM)** or **Topics Over Time (ToT)**, which account for the temporal dimension of the news data. These models could provide a more nuanced understanding of how peacebuilding narratives evolve over time. Additionally, **Correlated Topic Models (CTM)** or **Structural Topic Models (STM)** could offer insights into the relationships between topics and clusters, which were not addressed by the LDA model used in this study. Cross-checking LDA findings with these advanced models could further validate the results and deepen the analysis.

Finally, the cross-cutting themes identified in this study—such as the influence of the Catholic Church, the role of the private sector, and the focus on domestic and international conflicts—demonstrate the diverse ways in which peacebuilding is represented in the media across different Latin American contexts. Future research should continue to expand on these findings, using interdisciplinary approaches and advanced computational methods to deepen our understanding of peacebuilding narratives and their impact on society.

Acknowledgments. This research was funded by the National Center for Historical Memory under contract number 80740-154-2021, and was administered by the Colombian Ministry of Science, Technology, and Innovation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boutros-Ghali, B.: An agenda for peace: preventive diplomacy, peacemaking and peace-keeping. Int. Relat. **11**, 201–218 (1992). https://doi.org/10.1177/004711789201100302
2. Barnett, M., Kim, H., O'Donnell, M., Sitea, L.: Peacebuilding: what is in a name? GG **13**, 35–58 (2007). https://doi.org/10.1163/19426720-01301004

3. Acosta, R.A., Brunet, M.A., Córdoba, J.C.: The quality of the journalistic information of elespectador.com. The case of the Plebiscite for Peace, 2016, in Colombia. Revista Latina de Comunicación Social (2017). [in Spanish]. https://doi.org/10.4185/RLCS-2017-1231
4. Wolfsfeld, G., Khouri, R., Peri, Y.: News about the other in Jordan and Israel: does peace make a difference? Polit. Commun. **19**, 189–210 (2002). https://doi.org/10.1080/10584600252907443
5. Tenenboim-Weinblatt, K., Hanitzsch, T., Nagar, R.: Beyond peace journalism: reclassifying conflict narratives in the Israeli news media. J. Peace Res. **53**, 151–165 (2016). https://doi.org/10.1177/0022343315609091
6. Pearce, J., Perea, C.M.: Post war and non war violences: learning about peace and peacebuilding from Latin America. Peacebuilding **7**, 247–253 (2019). https://doi.org/10.1080/21647259.2019.1632057
7. Jacobi, C., van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit. J. **4**, 89–106 (2016). https://doi.org/10.1080/21670811.2015.1093271
8. Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R.: How to analyze political attention with minimal assumptions and costs. Am. J. Polit. Sci. **54**, 209–228 (2010). https://doi.org/10.1111/j.1540-5907.2009.00427.x
9. Neuendorf, K.A., Kumar, A.: Content analysis. In: Mazzoleni, G. (ed.) The International Encyclopedia of Political Communication, pp. 1–10. Wiley (2016). https://doi.org/10.1002/9781118541555.wbiepc065
10. Newman, D.J., Block, S.: Probabilistic topic decomposition of an eighteenth-century American newspaper. J. Am. Soc. Inf. Sci. **57**, 753–767 (2006). https://doi.org/10.1002/asi.20342
11. Maier, D., et al.: Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun. Methods Meas. **12**, 93–118 (2018). https://doi.org/10.1080/19312458.2018.1430754
12. Bird, S., Klein, E., Loper, E.: Natural language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc., Newton (2009)
13. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
14. DiMaggio, P., Nag, M., Blei, D.: Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. government arts funding. Poetics **41**, 570–606 (2013). https://doi.org/10.1016/j.poetic.2013.08.004
15. Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., Tolonen, M.: Topic modelling discourse dynamics in historical newspapers (2021). https://doi.org/10.5281/ZENODO.5562340
16. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta (2010)
17. Röder, M., Both, A., Hinneburg, A.: Exploring the Space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408. ACM, Shanghai (2015). https://doi.org/10.1145/2684822.2685324
18. Sievert, C., Shirley, K.: LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70. Association for Computational Linguistics, Baltimore (2014). https://doi.org/10.3115/v1/W14-3110
19. García-Perdomo, V., Harlow, S., Brown, D.K.: Framing the Colombian peace process: between peace and war journalism. Journal. Pract., 1–24 (2022). https://doi.org/10.1080/17512786.2022.2062428

A Comparative Study of Nonlinear PI and PI-PBC Control Strategies for Single-Phase PWM-CSC Using Adaptive Load Estimation

Angélica Mercedes Nivia-Vargas[1,2](✉), Oscar Danilo Montoya[1,2], and Walter Gil-Gonzalez[2,3]

[1] Universidad Distrital Francisco José de Caldas, Doctoral student in Engineering, Bogotá, Colombia
{amniviav,odmontoyag}@udistrital.edu.co

[2] Grupo de Compatibilidad e Interferencia Electromagnética (GCEM), Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C. 110121, Colombia

[3] Department of Electrical Engineering, Universidad Tecnológica de Pereira, Pereira, Colombia
wjgil@udistrital.edu.co

Abstract. This comparative study examines the application of nonlinear control strategies to pulse width modulation-controlled current source converters (PWM-CSCs), with a focus on proportional-integral (PI) control and proportional-integral passivity-based control (PI-PBC). PWM-CSCs play a crucial role in power electronics, enabling precise current flow regulation and control, which is essential for maintaining stability and achieving accurate management in electrical systems. Additionally, this paper highlights the significance of adaptive load estimation, which facilitates real-time adjustments to control strategies, thereby enhancing system efficiency and performance. By thoroughly analyzing dynamic models and control methodologies, this study demonstrates how the PI-PBC approach effectively stabilizes the system's incremental model to ensure optimal operation. The findings contribute to the advancement of control techniques in power electronics, offering valuable insights into the implementation of strategies for optimizing current regulation in complex electrical systems.

Keywords: Non-Linear Control · System Stability · Dynamic Modeling · Current Regulation

1 Introduction

1.1 General Context and Motivation

In power electronics, pulse width modulation-controlled current source converters (PWM-CSCs) have become essential devices for precise current regulation in

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 108–124, 2025.
https://doi.org/10.1007/978-3-031-88854-0_8

a variety of electrical systems. Unlike voltage source converters (VSCs), which primarily focus on voltage control, PWM-CSCs are specifically designed for current regulation [9]. This design makes them particularly suitable for applications requiring precise current control and robust system stability. Such converters are widely used in renewable energy systems, motor drives, and grid-connected applications, where accurate current control is vital for a reliable and efficient operation [27]. The ability of PWM-CSCs to maintain stable current under diverse operating conditions underscores their increasing importance in modern power systems, especially as renewable energy sources and distributed generation become more prevalent.

The growing complexity of power systems and the evolving nature of electrical loads have led to an increasing demand for sophisticated PWM-CSC control strategies. The inherent nonlinearities of these devices pose significant challenges to conventional control methods, particularly when faced with varying load conditions or rapidly changing system dynamics [19]. To address these issues, nonlinear control techniques, such as passivity-based control (PBC), have gained prominence as robust alternatives capable of handling complex dynamics [10].

This paper presents a comparative study of nonlinear control strategies, focusing on proportional-integral (PI) control and PI-PBC in the context of single-phase PWM-CSCs. The performance of these control methods is analyzed under various operating conditions, particularly against load variations, evaluating key behavioral parameters such as steady-state performance and harmonic content.

A significant element of this study is the integration of adaptive load estimation into control strategies. Adaptive load estimation is critical for PWM-CSCs [5], as it enables real-time adjustments to control parameters in response to fluctuating load demands, thus enhancing system performance [23]. This adaptability ensures efficient converter operation by minimizing energy waste and improving overall system stability [3]. The ability of control systems to dynamically respond to load variations is particularly vital in modern power systems, where loads are often unpredictable and may fluctuate rapidly [4]. By incorporating adaptive load estimation with nonlinear control strategies, this study evaluates the impact of real-time load adjustments on the effectiveness of both PI and PI-PBC controllers.

1.2 Fundamental Concepts

1.2.1 Pulse Width Modulation Current Source Converters

PWM-CSCs are widely utilized in power electronic applications given their capacity for current regulation within electrical systems [14]. These converters operate by modulating the duty cycle of switching elements, thereby regulating the current injected into or drawn from a load. In single-phase systems, PWM-CSCs are especially effective in applications requiring precise current control, such as renewable energy integration and high-performance motor operation [18]. The primary advantage of PWM-CSCs over VSCs lies in their direct current regulation capabilities, making them well-suited for scenarios where current stability is more critical than voltage regulation [14]. Maintaining stable and well-regulated current in single-phase applications is challenging due to the inherent nonlinear behavior of converters and the dynamic nature of loads [15]. This nonlinearity, coupled with the need for a rapid response to changes in operating conditions, requires advanced control strategies capable of addressing these complexities while ensuring a stable operation [21]. While traditional linear control techniques may suffice for simpler systems, they often fall short in more demanding situations, thus necessitating the use of nonlinear control methods.

1.2.2 Nonlinear Proportional-Integral Control

The nonlinear PI approach extends the traditional PI control strategy to address the limitations of linear PI controllers in systems with significant nonlinearities [33]. Although conventional PI control has been widely adopted in power electronics due to its simplicity and ability to correct steady-state errors, its performance deteriorates in the presence of system nonlinearities and time-varying dynamics, as encountered in PWM-CSCs [32]. Nonlinear PI control dynamically adjusts proportional and integral gains by considering the system's nonlinear characteristics, thereby enhancing robustness [28]. For PWM-CSCs, *i.e.*, inherently nonlinear systems, nonlinear PI control modifies the modulation index not only based on the reference current, but also in response to the system's nonlinearities [22]. This approach helps to mitigate the overshoot, oscillations, and slow response times implied under rapidly changing load conditions or when operating far from the nominal points. Consequently, nonlinear PI control offers improved stability and performance, making it more suitable for current regulation in complex power electronic systems [29,34].

1.2.3 PI Passivity-Based Control

The PI-PBC approach is a nonlinear control strategy that leverages the passive properties of a system to achieve stability [13]. Specifically, PI-PBC incorporates proportional-integral actions within the passivity framework. This approach uses the port-Hamiltonian representation of the system, where dynamics are defined in terms of energy storage, dissipation, and interconnection [30]. The port-Hamiltonian structure is particularly appropriate for PWM-CSCs, as it effectively captures the converter's bilinear dynamics. PI-PBC guarantees global asymptotic stability through the design of a control law based on the system's passive behavior, ensuring energy dissipation over time and leading to a stable equilibrium [11]. PI-PBC effectively controls current regulation by explicitly accounting for system nonlinearities, especially under varying load conditions and disturbances [31].

1.2.4 Adaptive Load Estimation

Adaptive load estimation offers precise load estimation by accounting for potential changes [7]. Along with responsiveness to fluctuations, it is essential for the optimal operation of PWM-CSCs [8]. Adaptive load estimation involves the real-time assessment of load conditions, with control parameters adjusted accordingly

[1]. This capability is crucial in applications with time-varying or unpredictable loads [36], such as renewable energy systems, where generation and consumption patterns can fluctuate rapidly. Incorporating adaptive load estimation into the control framework allows PI and PI-PBC controllers to respond effectively to changing operation environments. For instance, the control strategy can adapt to sudden load increases by modifying the modulation index or the control gains to maintain a stable and efficient operation. This adaptability improves control performance and enhances the overall system reliability, reducing the risks of instability and excess energy consumption [2].

1.3 Research Objective

The main objective of this study is to evaluate the performance of nonlinear PI control and PI-PBC strategies for single-phase PWM-CSCs, with a specific emphasis on the role of adaptive load estimation [35]. By examining how these control strategies address varying load conditions and system nonlinearities, this work provides a comprehensive analysis of their effectiveness in managing current regulation and ensuring system stability. Through a detailed exploration of the underlying dynamic models, control formulations, and stability criteria, this research advances the development of advanced control methodologies in the field of power electronics. Ultimately, this study aims to pave the way for more efficient, reliable, and adaptable current-controlled power systems.

1.4 Document Structure

This paper begins with an introduction to PWM-CSCs, emphasizing the importance of nonlinear control strategies, particularly PI control and PI-PBC. The subsequent section presents the dynamic model of the studied PWM-CSC, which serves as the basis for the analysis. The third section outlines the methodology employed to evaluate the efficacy of the control strategies, while the fourth one concentrates on the load estimation process, which is of paramount importance in enabling adaptive control. In the fifth section, the experimental results and the simulations are presented, showcasing the performance of both controllers under various load conditions. In the conclusion, potential avenues for future research on control systems for power electronics are outlined.

2 Dynamical Model

2.1 PWM-CSC Dynamics

The electrical topology of the studied PWM-CSC is depicted in Fig. 1. This configuration allows for independent current control, regardless of the input voltage, through electronically controlled switches such as IGBT or MOSFET transistors. This setup enables the modulation of both the magnitude and the direction of the output current [37]. The PWM strategy facilitates the generation of control

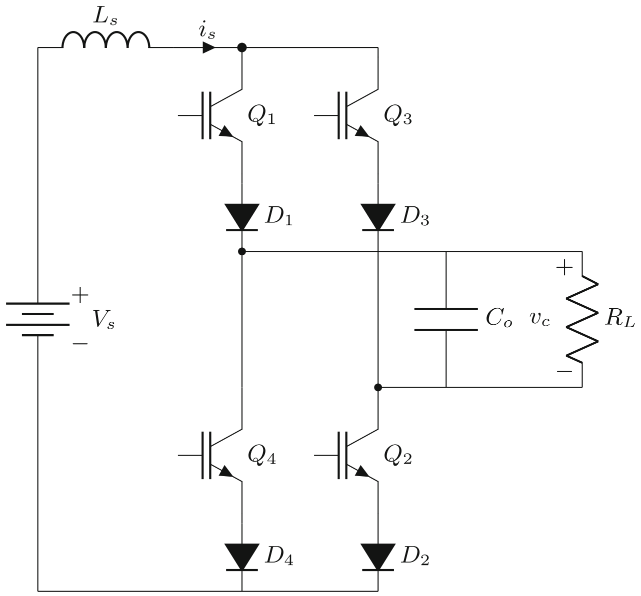


Fig. 1. Configuration PWM-CSC

signals for the inverter switches based on the computed modulation index, which is represented by m.

The PWM-CSC's configuration includes an output filter that absorbs the current generated by harmonics, utilizing the capacitor C_o. Consequently, the system can be modeled using state variables corresponding to the current in the inductor (i_s) and the voltage across the capacitor (v_c). By applying Kirchhoff's laws, Eqs. (1) and (2) are formulated.

$$L_s \frac{di_s(t)}{dt} = V_s(t) - R_s i_s(t) - m v_c(t), \tag{1}$$

$$C_o \frac{dv_c(t)}{dt} = m i_s(t) - G_L v_c(t). \tag{2}$$

The DC input source is represented by V_s, and L_s is a boosting inductor designed to absorb the harmonics generated during PWM switching. This inductor is modeled with a series resistance R_s. Additionally, G_L represents the equivalent conductance of a linear load, which is expressed as $G_L = \frac{1}{R_L}$. In other words, when expressed in terms of the state variables $x_1(t)$ and $x_2(t)$, the system is characterized by

$$L_s \dot{x}_1(t) = V_s(t) - R_s x_1(t) - u(t) x_2(t), \tag{3}$$

$$C_o \dot{x}_2(t) = u(t) x_1(t) - G_L x_2(t). \tag{4}$$

It should be noted that only two switches can operate at any given time. The PWM signal, which determines whether the transistors are activated or not, is the result of a comparison between a reference signal and the triangular carrier signal, which has been adjusted to have a frequency of $10kHz$. Thus, to achieve the desired operation, $x_2^*(t) = V_p \sin(\omega t + \theta)$, where V_p is the peak voltage, ω is the angular frequency, and θ represents the phase angle.

Remark 1. The nonlinear control design of the PWM-CSC modeled by Eqs. (3) and (4) requires determining the admissible trajectory for all the state variables. To this effect, the desired operating point (x_1^*, x_2^*) must be evaluated in the dynamical model, which yields the following desired state-space dynamics:

$$L_s \dot{x}_1^*(t) = V_s(t) - R_s x_1^*(t) - u(t)^* x_2^*(t), \tag{5}$$

$$C_o \dot{x}_2^*(t) = u(t)^* x_1^*(t) - G_L x_2^*(t), \tag{6}$$

where the control input at the equilibrium point (*i.e.*, $u(t)^*$) can be obtained from both equations. However, when using Eq. (5), a division by $x_2^*(t)$ will appear, so the best option to obtain this control input is through (6).

$$u(t)^* = \frac{1}{x_1^*(t)} \left(C_o \dot{x}_2^*(t) + G_L x_2^*(t) \right). \tag{7}$$

Note that the main goal of the PWM-CSC application studied in this research is to provide a sinusoidal voltage profile to a linear load (see Fig. 1). The value of the desired variable $x_2^\star$ is fully known, *i.e.*, $x_2^\star = V_m \sin(\omega t)$, where V_m is the voltage desired amplitude, and ω represents the desired angular frequency.

3 Control Strategies

3.1 Nonlinear PI Control

The objective of the proposed Nonlinear PI controller is to ensure the asymptotic convergence of x_2 and to maintain a tight behavior of x_1 for all $t \geq 0$ [18]. To this effect, it is assumed that the function following the reference is given by:

$$x_1(t) = I_s + I_h \sin(2\omega t + \varphi). \tag{8}$$

The DC value of i_s is represented by I_s, while the frequency should be twice that of the AC side. The amplitude of the harmonic ripple, denoted by I_h, is given by:

$$I_h = \sqrt{\left(\frac{V_m^2}{4\omega L I_s R_L} \right)^2 + \left(\frac{C V_m^2}{4\, L I_s} \right)^2}, \tag{9}$$

and the phase is as follows:

$$\varphi = \tan^{-1}(\omega C R_L). \tag{10}$$

Considering that it is defined as $u(t) = u^*(t) + \tilde{u}(t)$, and following the recommendations of [25], this control input takes the following form:

$$u(t) = \frac{1}{x_1^*(t)} \left(C_o \dot{x}_2^*(t) + G_L x_2^*(t) - k_p \tilde{x}_2 - k_i \int_0^t \tilde{x}_2 \right), \tag{11}$$

where k_p and k_i are constant values representing the proportional and integral action, and $\tilde{x}_2 = x_2 - x_2^*$ is intended to be zero.

Remark 2. The authors of [17] recommended avoiding the division by $x_1^*(t)$ in the Control Input (11), as it would introduce some oscillations in the average control input. These oscillations can be easily absorbed by the PI control component, which implies that, for the sake of simplicity in the control implementation, the following control input is implemented:

$$u(t) = \frac{1}{I_s} \left(C_o \dot{x}_2^*(t) + G_L x_2^*(t) - k_p \tilde{x}_2 - k_i \int_0^t \tilde{x}_2 \right). \tag{12}$$

The Lyapunov stability theorem allows ascertaining the stability of a linear or nonlinear system [16]. This is done by examining the existence of a positive-definite Lyapunov function $V(x)$, for which the time derivative $\dot{V}(x)$ is negative-definite. The negative-definiteness of $\dot{V}(x)$ implies that the equilibrium point at the origin $(x_1,\, x_2)$ is globally asymptotically stable. Consequently, the control strategy proposed herein ensures that x_2 is asymptotically stable. By substituting the Eq. (12) into (4), the following is obtained:

$$C_o \dot{x}_2(t) = -(k_p + R_L^{-1}) \tilde{x}_2 - k_i \int_0^t \tilde{x}_2. \tag{13}$$

This implies that $\tilde{x}_2$ tends to zero, thereby ensuring that $x_2 \rightarrow x_2^*$, which guarantees an asymptotic behavior towards the reference. However, in the desired Control Input Reference (7), it can be observed that knowledge of G_L is required to ensure a reliable supply of real-time load demand information, which can then be used to adjust the control action. This estimation, in turn, contributes to a more efficient inverter operation, thereby avoiding unnecessary energy supply. Therefore, a load estimator design is proposed, as described in the next section.

3.2 PI-PBC

The proportional-integral passivity-based control (PI-PBC) framework is a well-established method in control theory, particularly suited for a specific class of nonlinear dynamical system characterized by a bilinear structure [6]. By utilizing classical PI controllers, this approach facilitates the design of controllers that are asymptotically stable (specifically in the sense of Lyapunov) by employing port-Hamiltonian theory in the context of passive systems [26].

In a recent work by [12], it was shown that the PI-PBC method is especially effective for controlling power electronic converters with linear loads. It manages to stabilize the converter at its equilibrium point by applying a PI controller to the system's passive output.

3.3 Port-Hamiltonian System

A PWM-CSC can be described as a port-Hamiltonian system, as expressed in the following structure:

$$\begin{bmatrix} L_s & 0 \\ 0 & C_o \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \left(\begin{bmatrix} 0 & -u(t) \\ u(t) & 0 \end{bmatrix} - \begin{bmatrix} R_s & 0 \\ 0 & G_L \end{bmatrix} \right) \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} V_s \\ 0 \end{bmatrix} \tag{14}$$

$$\mathcal{D}\dot{x} = \left(\mathcal{J}\left(u\right) - \mathcal{R}\right) x + \vartheta, \tag{15}$$

In this representation, $\mathcal{D}$ refers to the energy storage matrix, and $\mathcal{J}\left(u\right)$ is the interconnection matrix, which is skew-symmetric, meaning that $\mathcal{J}\left(u\right) + \mathcal{J}^{\top}\left(u\right) = 0$. Additionally, $\mathcal{R}$ represents the dissipation matrix, and ϑ is the vector of external inputs. The specific structure of each matrix can be derived by comparing Eqs. (14) and (15).

Remark 3. The design of a PI-PBC controller necessitates the definition of an incremental model, which requires the existence of an equilibrium trajectory or an operating point. In the case of the PWM-CSC, an equilibrium trajectory is indeed present.

The dynamical system expressed in (15) admits an equilibrium trajectory, which denoted by $x^{\star}$. The corresponding dynamic behavior is described by the following equation:

$$\mathcal{D}\dot{x}^{\star} = \left(\mathcal{J}\left(u^{\star}\right) - \mathcal{R}\right) x^{\star} + \vartheta, \tag{16}$$

where $\dot{x}^{\star}$ represents the time derivative of the trajectory (which, for single-phase converter applications, is nonzero), and $u^{\star}$ represents a smooth control input.

To obtain the incremental model, which characterizes the system's error dynamics, the state variable errors are defined as $\tilde{x} = x - x^{\star}$ and $\tilde{u} = u - u^{\star}$. By subtracting the desired Dynamical System (16) from (15) and performing the necessary algebraic manipulations, the incremental model is obtained, as shown in Eq. (17).

$$\mathcal{D}\dot{\tilde{x}} = \mathcal{J}\left(\tilde{u}\right) x^{\star} + \left(\mathcal{J}\left(u\right) - \mathcal{R}\right) \tilde{x}, \tag{17}$$

Theorem 1 (PI-PBC law). *The incremental model indicated in* (17) *can be asymptotically stabilized by defining an appropriate PI control law, as outlined by [25], leading the system to a stable equilibrium at* $\tilde{x} = 0$ *when the control input* $\tilde{u}$ *is given by Eq.* (18).

$$\begin{aligned} \tilde{u} &= -k_p \tilde{y} - k_i \tilde{z}, \\ \dot{\tilde{z}} &= \tilde{y}, \end{aligned} \tag{18}$$

where k_p *and* k_i *are positive control gains, and* $\tilde{y}$ *represents the passive output of the system.*

Proof. To determine whether the control law defined in (18) stabilizes the Incremental Model (17), the principles of Lyapunov stability are employed.

Theorem 2 (Lyapunov stability theorem). *A dynamical system* $\dot{\tilde{x}} = f(\tilde{x})$ *is asymptotically stable at the origin* ($\tilde{x} = 0$) *if there is a Lyapunov candidate function* $\mathcal{V}(\tilde{x})$ *that satisfies the following conditions:*

1. $\mathcal{V}(\tilde{x})$ *is positive-definite, i.e.,* $\mathcal{V}(\tilde{x}) > 0$ *for all* $\tilde{x} \neq 0$, *and* $\mathcal{V}(0) = 0$.
2. *The time derivative* $\dot{\mathcal{V}}(\tilde{x})$ *is negative-definite, i.e.,* $\dot{\mathcal{V}}(\tilde{x}) < 0$.

Based on Theorem 2, a candidate Lyapunov function is defined in (19), whose time derivative is calculated in (20).

$$\mathcal{V}(\tilde{x}, \tilde{z}) = \frac{1}{2}\tilde{x}^\top \mathcal{D}\tilde{x} + \frac{1}{2}k_i\tilde{z}^2, \tag{19}$$

$$\dot{\mathcal{V}}(\tilde{x}, \tilde{z}) = \tilde{x}^\top \mathcal{D}\dot{\tilde{x}} + k_i\tilde{z}\dot{\tilde{z}}. \tag{20}$$

Substituting the Incremental Model (17) into this Lyapunov function results in (21).

$$\begin{aligned}\dot{\mathcal{V}}(\tilde{x}, \tilde{z}) &= \tilde{x}^\top \left(\mathcal{J}(\tilde{u})\, x^\star + \left(\mathcal{J}(u) - \mathcal{R}\right)\tilde{x}\right) + k_i\tilde{z}\tilde{y} \\ &= -\tilde{x}^\top \mathcal{R}\tilde{x} + \tilde{x}^\top \mathcal{J}(\tilde{u})\, x^\star + k_i\tilde{z}\tilde{y}\end{aligned} \tag{21}$$

By exploiting the properties of the skew-symmetric matrix, it can be sated that $\tilde{x}^\top \mathcal{J}(\tilde{u})\,\tilde{x} = 0$ [20]. Additionally, given the structure of the interconnection matrix, the following holds:

$$\mathcal{J}(\tilde{u})\, x^\star = \mathcal{J}_1 x^\star \tilde{u} \implies \mathcal{J}_1 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Substituting this result into Eq. (21) along with the control input defined in (18) leads to the final stabilization result *i.e.*, (23).

$$\dot{\mathcal{V}}(\tilde{x}, \tilde{z}) = -\tilde{x}^\top \mathcal{R}\tilde{x} + \tilde{x}^\top \mathcal{J}_1 x^\star \tilde{u} + k_i\tilde{z}\tilde{y} \tag{22}$$

$$= -\tilde{x}^\top \mathcal{R}\tilde{x} + \tilde{x}^\top \mathcal{J}_1 x^\star \left(-k_p\tilde{y} - k_i\tilde{z}\right) + k_i\tilde{z}\tilde{y}, \tag{23}$$

A port-Hamiltonian system with the structure defined by (18) has a passive output, as characterized in Eq. (24).

$$\tilde{y}^\top = \tilde{x}^\top \mathcal{J}_1 x^\star \implies \tilde{y} = x^{\star,\top} \mathcal{J}_1 \tilde{x}. \tag{24}$$

Finally, considering the passive output definition in (24), the time derivative of the Lyapunov candidate function is reduced to

$$\begin{aligned} \dot{\mathcal{V}}(\tilde{x}, \tilde{z}) &= -\tilde{x}^\top \mathcal{R}\tilde{x} - \tilde{y}^\top k_p \tilde{y} - k_i \tilde{y}^\top \tilde{z} + k_i \tilde{z}\tilde{y}, \\ &= -\tilde{x}^\top \mathcal{R}\tilde{x} - \tilde{y}^\top k_p \tilde{y} \leq 0, \end{aligned} \tag{25}$$

This confirms that the PI-PBC control input defined in (18) successfully stabilizes the incremental model at the origin, thus completing the proof. The asymptotic stability of the system is further validated by applying LaSalle's invariance principle, as demonstrated by [6].

To complete the controller, *i.e.*, $u = \tilde{u} + u^\star$, the passive output $\tilde{y}$ is calculated, which yields

$$\tilde{y} = x^{\star,\top} \mathcal{J}_1 \tilde{x} = \begin{bmatrix} x_1^\star & x_2^\star \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^\top \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} \implies \tilde{y} = x_1^\star \tilde{x}_2 - x_2^\star \tilde{x}_1. \tag{26}$$

This produces the following PI-PBC controller:

$$u = \frac{1}{x_1^\star}\left(C_o \dot{x}_2^\star + G_L x_2^\star\right) - k_p \left(x_1^\star \tilde{x}_2 - x_2^\star \tilde{x}_1\right) - k_i \int_0^{t_f} \left(x_1^\star \tilde{x}_2 - x_2^\star \tilde{x}_1\right) d\tau. \tag{27}$$

Remark 4. To implement the PI-PBC control law defined by (27) in a PWM-CSC application, it is necessary to calculate the trajectory followed by the inductor current, *i.e.*, $x_1^\star$. This trajectory is reached by solving the differential equation of the desired dynamical system in (17), which is presented below:

$$L_s \dot{x}_1^\star = V_s - R_s x_1^\star - u^\star x_2^\star. \tag{28}$$

4 Load Estimation

Considering the information provided by the state variables and the model described in Eqs. (3) and (4), an integral estimator is proposed, as shown in Eq. (33). This estimator is specifically crafted to achieve asymptotic convergence to the load value, employing the corresponding conductance G_L [24].

$$\widehat{G}_L = G_{LI} + \alpha f(x_2), \tag{29}$$

The goal of the estimation process, represented by $\widehat{G}_L$, is to ensure convergence towards G_L as the time t approaches infinity. To facilitate this convergence, an initial estimate is defined: $\tilde{G}_L(t) = \tilde{G}_L(0)e^{-\alpha\beta t}$.

The estimation error is defined as

$$\tilde{G}_L = \widehat{G}_L - G_L. \tag{30}$$

In designing the estimator, it is important that $\tilde{G}_L \to 0$ as $t \to \infty$. Consequently, the time derivative of the estimation error, as indicated in Eq. (30), can be expressed as

$$\dot{\tilde{G}}_L = \dot{G}_{LI} - G_L + \alpha \dot{f}(x_2)\dot{x}_2. \tag{31}$$

From this, the formulation for $\dot{G}_{LI}$ is derived.

$$\dot{G}_{LI} = \alpha \dot{f}(x_2)\left(\frac{u(t)}{C}x_1 - \frac{\tilde{G}_L}{C}x_2\right) + \alpha(x_2^2 - \beta)\widehat{G}_L. \tag{32}$$

Thus, the estimation is articulated as

$$\widehat{G}_L = G_{LI} + \alpha f(x_2). \tag{33}$$

The parameters and functions G_{LI}, α, and $f(x_2)$ are specified in accordance with Eqs. (34) and (32). The estimator requires inputs including u, the control signal, and the state measurements x_1 and x_2, as depicted in Fig. 2.

$$f(x_2) = -\frac{1}{2}C_o x_2^2. \tag{34}$$

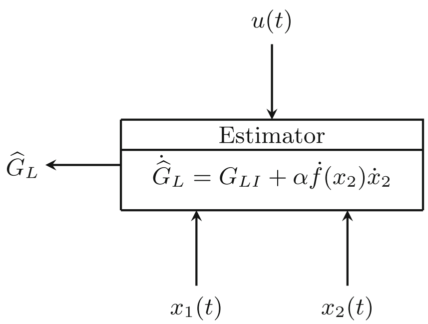


Fig. 2. Load estimator

5 Simulations and Results

The proposed controller design was implemented in a simulated environment using the Matlab ®software. The performance results of the proposed controllers and load estimator were analyzed. To conduct the simulation, the following parameter values were defined, with the aim of interpreting the results obtained: $V_p = 150$V, $V_s = 48$V, $L_s = 10$mH, $C_o = 200\mu$F, and $R_s = 1\Omega$. These values were selected according to [17].

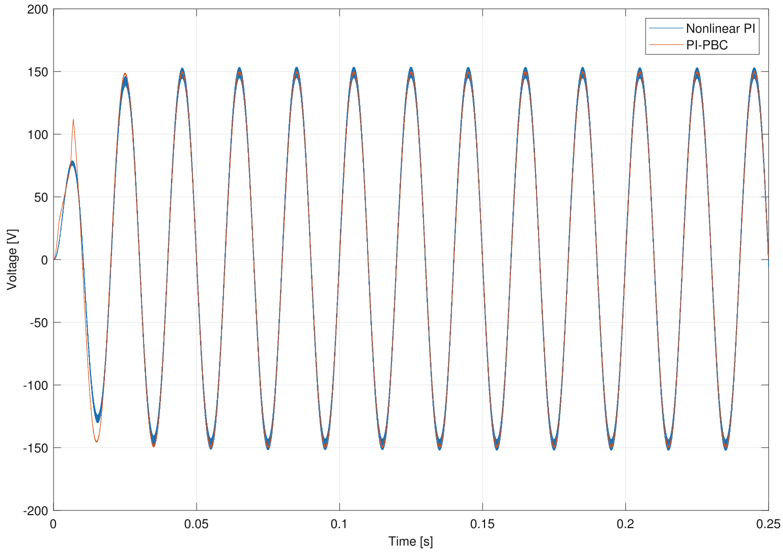


Fig. 3. Voltage reference response

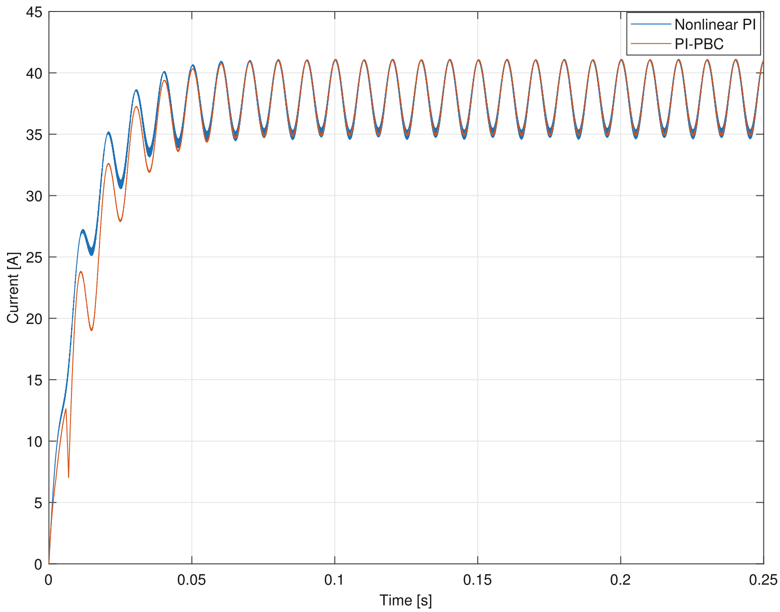


Fig. 4. Following the current reference signal

First, an experimental tuning of the PI controller constants was conducted with $k_p = 0.2$ and $k_i = 0.2$, allowing the pulse-width modulation system to adjust the switching and achieve the desired behavior $x_2^*(t) = V_p \sin(\omega t + \theta)$V. On the other hand, the PI-PBC constants were set to $K_p = 1.5$ and $K_i = 0.8$. When the implementation was performed with a constant load of 50Ω, the resulting voltage behavior was observed, as detailed in Fig. 3 for both controllers. While both strategies reached the desired value, the PI-PBC controller exhibited a superior performance by reducing the response variation. Furthermore, the steady-state response achieved the expected value and the desired behavior.

Likewise, the response demonstrates a notable improvement in behavior, exhibiting a reduction of approximately 5% in the steady-state error. However, the PI controller attains this steady state more quickly, as evidenced in Fig. 4.

Table 1. Comparison of the control parameters

Parameter	PI-PBC	Nonlinear PI
V_p	149.4	150.3 V
V_{rms}	105.6	106.3 V
I_p	2.986 A	3.006 A
I_{rms}	2.112 A	2.125 A
THD_I	3.35%	2.72%
Steady-state error max	0.05 A	0.26 A
Settling Time	56 ms	47 ms

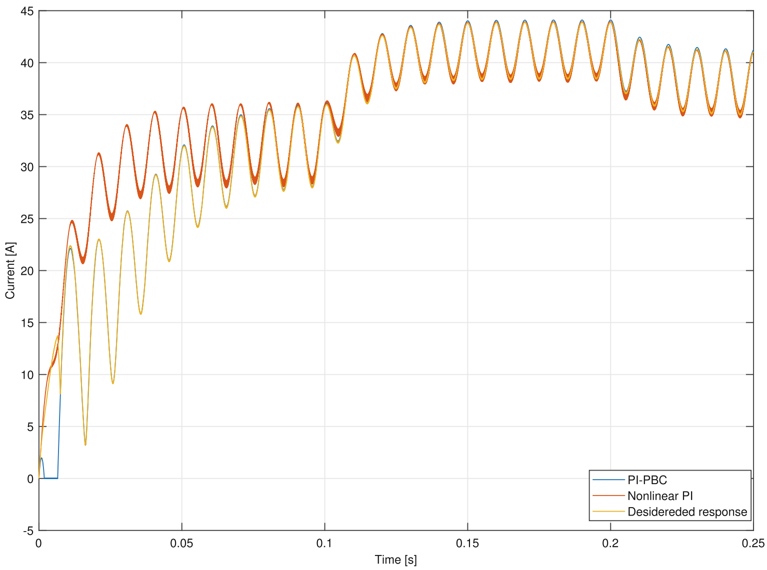


Fig. 5. Tracking the current under variable load conditions

Table 1 presents a comparison of the performance parameters for each case.

In the case involving a variable load that is combined with the results of the estimator, it is worth highlighting that the PI controller reaches the desired value more rapidly. However, once the steady state is achieved, and in the face of a load variation, the PI-PBC responds with a precise tracking of the changes in current, as shown in Fig. 5.

Likewise, it can be observed that, as expected, the voltage behavior is always guaranteed to be at the desired value, regardless of the load demand. This is evidenced in Fig. 6.

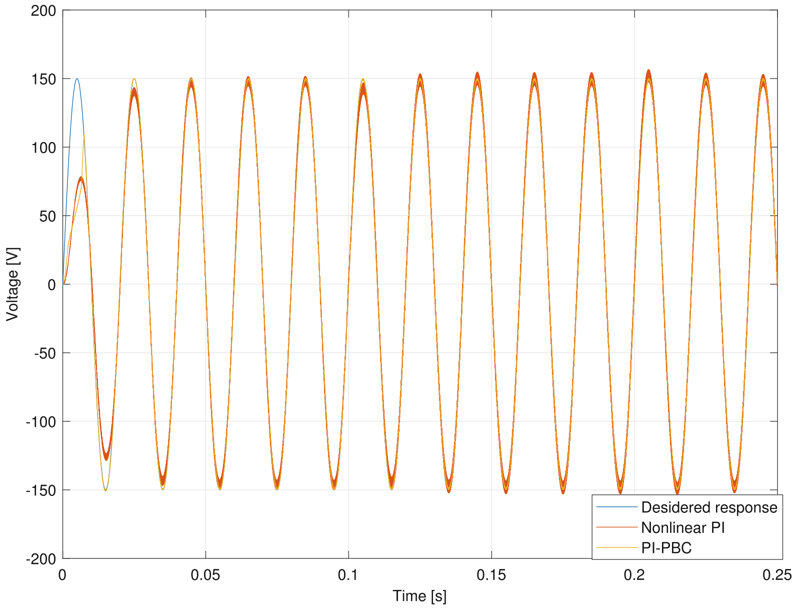


Fig. 6. Tracking the voltage under variable load conditions

6 Conclusions

The results of this study indicate that both the PI and PI-PBC controllers are capable of achieving the desired current levels. However, the PI-PBC demonstrates a superior performance profile. Specifically, it reduces the response variation and steady-state error compared to the traditional PI controller. This enhanced stability is necessary for applications that require precise current regulation, highlighting the effectiveness of PI-PBC in managing the complexities associated with nonlinear systems.

While the PI controller is faster in reaching steady-state conditions, the PI-PBC stands out due to its impressive capability to accurately track changes in load once the steady state is attained. This distinction suggests that the PI-PBC is better at handling dynamic environments with fluctuating load conditions. Consequently, the PI-PBC's ability to maintain its performance under variable loads makes it a preferable choice in applications that demand adaptability and reliability.

In the context of this study, parameter tuning was pivotal in achieving the desired system behavior. The PI controller was tuned with $k_p = 0.2$ and $k_i = 0.2$, whereas the PI-PBC utilized $k_p = 1.5$ and $k_i = 0.8$. These parameter selections were instrumental in enhancing the overall performance of the pulse-width modulation (PWM) system, and the higher gains of the PI-PBC allowed it to effectively deal with the inherent nonlinearities of the PWM-controlled systems.

The PI-PBC demonstrated a superior capacity for adapting to variable load conditions, which is particularly significant in dynamic electrical systems. Its performance underscores the usefulness of this control strategy in real-world applications where load changes are frequent and unpredictable. By maintaining a more consistent current regulation in these scenarios, the PI-PBC constitutes an invaluable asset in the design and implementation of advanced power electronic systems.

Ultimately, the findings of this research contribute significantly to the ongoing development of control techniques within the field of power electronics. By providing valuable insights into the implementation of control strategies to optimize current regulation in complex electrical systems, this study lays the groundwork for future advancements. In conclusion, while both controllers are effective, the PI-PBC offers notable advantages in terms of stability and adaptability, particularly in environments characterized by fluctuating load conditions.

References

1. Abdel-Rady Ibrahim Mohamed, Y., El-Saadany, E.F.: An improved deadbeat current control scheme with a novel adaptive self-tuning load model for a three-phase PWM voltage-source inverter. IEEE Trans. Ind. Electron. **54**(2), 747–759 (2007). https://doi.org/10.1109/tie.2007.891767
2. Albogamy, F.R., et al.: An optimal adaptive control strategy for energy balancing in smart microgrid using dynamic pricing. IEEE Access **10**, 37396–37411 (2022). https://doi.org/10.1109/access.2022.3164809
3. Babu, T.S., Vasudevan, K.R., Ramachandaramurthy, V.K., Sani, S.B., Chemud, S., Lajim, R.M.: A comprehensive review of hybrid energy storage systems: converter topologies, control strategies and future prospects. IEEE Access **8**, 148702–148721 (2020). https://doi.org/10.1109/access.2020.3015919
4. Callaway, D.S., Hiskens, I.A.: Achieving controllability of electric loads. Proc. IEEE **99**(1), 184–199 (2011). https://doi.org/10.1109/jproc.2010.2081652
5. Chavarro-Barrera, L., Perez-Londono, S., Mora-Florez, J.: An adaptive approach for dynamic load modeling in microgrids. IEEE Trans. Smart Grid **12**(4), 2834–2843 (2021). https://doi.org/10.1109/tsg.2021.3064046
6. Cisneros, R., Pirro, M., Bergna, G., Ortega, R., Ippoliti, G., Molinas, M.: Global tracking passivity-based PI control of bilinear systems: application to the interleaved boost and modular multilevel converters. Control. Eng. Pract. **43**, 109–119 (2015). https://doi.org/10.1016/j.conengprac.2015.07.002

7. Crocker, S.J., Mathieu, J.L.: Adaptive state estimation and control of thermostatic loads for real-time energy balancing. In: 2016 American Control Conference (ACC), pp. 3557–3563. IEEE (2016). https://doi.org/10.1109/acc.2016.7525465
8. Danilo, O., Giraldo, M.: Passivity-based analysis and control of ac microgrids: integration, operation and control of energy storage systems (2019). https://hdl.handle.net/11059/10132
9. Ding, L., Li, Y.: High-power current source converters, pp. 367–402. Elsevier (2021). https://doi.org/10.1016/b978-0-12-819432-4.00003-2
10. Farooq, M.U., Eizad, A., Bae, H.K.: Power solutions for autonomous mobile robots: a survey. Rob. Auton. Syst. **159**, 104285 (2023). https://doi.org/10.1016/J.ROBOT.2022.104285
11. Gil-Gonzalez, W., Garces, A., Montoya, O.D.: Current PI control for PV systems in DC microgrids: a PBC design, vol. 2014, pp. 1–5 (2019). https://doi.org/10.1109/pepqa.2019.8851555
12. Gil-González, W., Montoya, O.D., Riffo, S., Restrepo, C., Muñoz, J.: A global tracking sensorless adaptive PI-PBC design for output voltage regulation in a boost converter feeding a DC microgrid. Energies **16**(3), 1106 (2023). https://doi.org/10.3390/en16031106
13. Gil-González, W., Montoya, O.D., Garces, A.: Direct power control of electrical energy storage systems: a passivity-based PI approach. Electr. Power Syst. Res. **175**, 105885 (2019). https://doi.org/10.1016/j.epsr.2019.105885
14. Giraldo, E., Garces, A.: An adaptive control strategy for a wind energy conversion system based on PWM-CSC and PMSG. IEEE Trans. Power Syst. **29**(3), 1446–1453 (2014). https://doi.org/10.1109/tpwrs.2013.2283804
15. Huang, X., et al.: Robust current control of grid-tied inverters for renewable energy integration under non-ideal grid conditions. IEEE Trans. Sustain. Energy **11**(1), 477–488 (2020). https://doi.org/10.1109/tste.2019.2895601
16. Kanakis, G.S., Rovithakis, G.A.: Guaranteeing global asymptotic stability and prescribed transient and steady-state attributes via uniting control. IEEE Trans. Autom. Control **65**(5), 1956–1968 (2020). https://doi.org/10.1109/tac.2019.2925504
17. Komurcugil, H.: Nonlinear control strategy for single-phase PWM current-source inverters. In: 2009 35th Annual Conference of IEEE Industrial Electronics. IEEE (2009). https://doi.org/10.1109/iecon.2009.5415002
18. Liu, P., Wang, Z., Wei, S., Bo, Y., Pu, S.: Recent developments of modulation and control for high-power current-source-converters fed electric machine systems. CES Trans. Electr. Mach. Syst. **4**(3), 215–226 (2020). https://doi.org/10.30941/cestems.2020.00027
19. Liu, Y.H., Watson, N.R., Zhou, K.L., Yang, B.F.: Converter system nonlinear modelling and control for transmission applications-part II: CSC systems. IEEE Trans. Power Delivery **28**(3), 1391–1401 (2013). https://doi.org/10.1109/tpwrd.2013.2250529
20. López-Rodríguez, K., Gil-González, W., Escobar-Mejía, A.: Design and implementation of a PI-PBC to manage bidirectional power flow in the dab of an SST. Res. Eng. **14**, 100437 (2022). https://doi.org/10.1016/j.rineng.2022.100437
21. Mahmoud, M.S., Alyazidi, N.M., Abouheaf, M.I.: Adaptive intelligent techniques for microgrid control systems: a survey. Int. J. Electr. Power Energy Syst. **90**, 292–305 (2017). https://doi.org/10.1016/j.ijepes.2017.02.008
22. Mahmud, M.A., Pota, H.R., Hossain, M.J.: Nonlinear current control scheme for a single-phase grid-connected photovoltaic system. IEEE Trans. Sustain. Energy **5**(1), 218–227 (2014). https://doi.org/10.1109/tste.2013.2279884

23. Mohsenian-Rad, A.H., Leon-Garcia, A.: Optimal residential load control with price prediction in real-time electricity pricing environments. IEEE Trans. Smart Grid **1**(2), 120–133 (2010). https://doi.org/10.1109/tsg.2010.2055903
24. Montoya, O.D., Serra, F.M., Gil-Gonzalez, W., Asensio, E.M., De Angelo, C.H.: Adaptive control of a single-phase grid-forming for feeding unknown resistive loads. In: 2021 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). IEEE (2021). https://doi.org/10.1109/ropec53248.2021.9668136
25. Montoya, O., Gil-Gonzalez, W., Garces, A., Serra, F., Hernandez, J.: PI-PBC approach for voltage regulation in Ćuk converters with adaptive load estimation. In: 2020 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). IEEE (2020). https://doi.org/10.1109/ropec50909.2020.9258716
26. Montoya, O.D., Serra, F.M., Espinosa-Pérez, G.: On the equivalence between PI-PBC and IOC designs: an application involving three-phase front-end converters. IEEE Trans. Circuits Syst. II Express Briefs **71**(1), 241–245 (2024). https://doi.org/10.1109/tcsii.2023.3299203
27. Mumtaz, F., Zaihar Yahaya, N., Tanzim Meraj, S., Singh, B., Kannan, R., Ibrahim, O.: Review on non-isolated DC-DC converters and their control techniques for renewable energy applications. Ain Shams Eng. J. **12**(4), 3747–3763 (2021). https://doi.org/10.1016/j.asej.2021.03.022
28. Ortega, R., Astolfi, A., Barabanov, N.E.: Nonlinear PI control of uncertain systems: an alternative to parameter adaptation. Syst. Control Lett. **47**(3), 259–278 (2002). https://doi.org/10.1016/s0167-6911(02)00212-8
29. Precup, R.E., David, R.C., Petriu, E.M., Radac, M.B., Preitl, S.: Adaptive GSA-based optimal tuning of pi controlled servo systems with reduced process parametric sensitivity, robust stability and controller robustness. IEEE Trans. Cybern. **44**(11), 1997–2009 (2014). https://doi.org/10.1109/tcyb.2014.2307257
30. Ramirez, H., Maschke, B., Sbarbaro, D.: Modelling and control of multi-energy systems: an irreversible port-hamiltonian approach. Eur. J. Control. **19**(6), 513–520 (2013). https://doi.org/10.1016/j.ejcon.2013.09.009
31. Riffo, S., Gil-González, W., Montoya, O.D., Restrepo, C., Muñoz, J.: Adaptive sensorless PI+passivity-based control of a boost converter supplying an unknown CPL. Mathematics **10**(22), 4321 (2022). https://doi.org/10.3390/math10224321
32. Sadigh, A.K., Dargahi, V., Corzine, K.: Analytical determination of conduction power loss and investigation of switching power loss for modified flying capacitor multicell converters. IET Power Electron. **9**(2), 175–187 (2016). https://doi.org/10.1049/iet-pel.2015.0369
33. Shi, L., Miao, X., Wang, H.: An improved nonlinear proportional-integral-differential controller combined with fractional operator and symbolic adaptation algorithm. Trans. Inst. Meas. Control. **42**(5), 927–941 (2019). https://doi.org/10.1177/0142331219879332
34. Song, Y., Wang, Y., Wen, C.: Adaptive fault-tolerant pi tracking control with guaranteed transient and steady-state performance. IEEE Trans. Autom. Control **62**(1), 481–487 (2017). https://doi.org/10.1109/tac.2016.2554362
35. Tõnso, M., Kaparin, V., Belikov, J.: Port-hamiltonian framework in power systems domain: a survey. Energy Rep. **10**, 2918–2930 (2023). https://doi.org/10.1016/j.egyr.2023.09.077
36. Wang, Q., Li, W., Mohajer, A.: Load-aware continuous-time optimization for multi-agent systems: toward dynamic resource allocation and real-time adaptability. Comput. Netw. **250**, 110526 (2024). https://doi.org/10.1016/j.comnet.2024.110526
37. Zhang, Y., Li, Y.: Current Source Converters and Their Control, pp. 115–140. Elsevier (2018). https://doi.org/10.1016/b978-0-12-816136-4.00016-6

A Linear Algebra Based Controller for an Autonomous Ground Vehicle: Commissioning and Testing

Gabriel Gómez-Guerra[1,2], René Játiva[1(✉)], and Gustavo Scaglia[3]

[1] Colegio de Ciencias e Ingenierías "El Politécnico", Universidad San Francisco de Quito USFQ, Quito, Ecuador
{ggomez,rjativa}@usfq.edu.ec

[2] Instituto de Energía y materiales, Universidad San Francisco de Quito USFQ, Quito, Ecuador

[3] Instituto de Ingeniería Química, Universidad Nacional de San Juan, San Juan, Argentina
gscaglia@unsj.edu.ar

Abstract. The Navigation Subsystem of an autonomous ground vehicle has been developed to follow a predefined trajectory along a semi-structured environment. For this purpose, a linear algebra-based controller has been designed taking into account the Ackerman kinematic model. Furthermore, two PID controllers and two one-dimensional Kalman filters have been implemented to allow the proper adjustment of the vehicle's linear velocity and steering angle. A CAN expansion module has also been installed in the vehicle Navigation Subsystem, enabling it to communicate with the vehicle Environment Description Subsystem, which provides information about the surroundings, allowing the Navigation Subsystem to react to external stimuli. Finally, several tests were performed to verify the correct behavior of the controllers and the communication interface.

Keywords: Linear algebra-based controller · Autonomous vehicle · Ackerman kinematic model · Kalman filter · CAN communications protocol · Trajectory tracking · PID controller

1 Introduction

An autonomous vehicle is defined as one capable of recognizing its environment and making decisions based on the information received. It can operate as a traditional vehicle without the intervention of a human user. However, the Society of Automotive Engineers (SAE) defines six levels of autonomy, ranging from fully manual to fully autonomous. Current commercial developments are classified between level two, where sensors are used to trigger alarms to assist the

This work has been partially funded with USFQ Poli-Grant 16931.

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 125–139, 2025.
https://doi.org/10.1007/978-3-031-88854-0_9

user, and level three, where control can be delegated to the vehicle, but the user needs to be alert to take control at any time [11]. The development of autonomous vehicles is gaining relevance due to the potential advantages they offer, such as the reduction of traffic accidents and smoother mobility [10,19].

Nowadays, the state of the art shows heterogeneous development, with small fleets delivering packages in low-traffic suburban areas, autonomous farm vehicles [14], and semi-autonomous services that can switch control between the driver and the vehicle according to the traffic scenario's complexity, allowing the driver to take control at any time, especially in emergency scenarios [10,11,19]. However, several applications can be implemented using autonomous ground vehicles designed to operate in semi-structured environments, such as package delivery within a university campus or as part of Industry 4.0 processes. Full autonomy is not required in these environments, where terrain and building architecture are previously known but dynamic obstacles appear.

This work involves the commissioning of a prototype vehicle developed to study and research autonomous ground vehicles and their applications [2] in semi-structured environments. It has been modeled as an Ackerman vehicle, where linear thrust is provided by the rear axle, while the direction is modified by the steering shaft at the front axle [1,6,7,15,16,20,22]. The mapping component has been excluded from the navigation system at this stage of the project, and the vehicle trajectory is provided as referenced routes to a coordinate system. Furthermore, a linear algebra-based controller is designed using the Ackerman kinematic model. This controller uses the vehicle's direction and its position in the reference coordinate system to calculate the vehicle's linear velocity and steering angle. Two PID controllers are implemented to allow the linear velocity and the steering angle to follow the reference generated by the main controller [6,7,16,20,22], and two one-dimensional Kalman filters are used to reduce the variance in measurements of the vehicle's linear velocity and steering angle.

Additionally, a Controller Area Network (CAN) expansion hat has been installed on the main Raspberry Pi processor to enable it to communicate with a secondary unit to obtain additional information about the surroundings, such as real-time image processing scanning, or results of the application of algorithms for robust mission planning [12]. The CAN protocol, originally developed for the automotive industry, allows multiple devices to share information across a bidirectional communication bus [8,9].

The remainder of this article is organized as follows: Sect. 2 describes the system, mathematical models and system design are treated in Sect. 3, results from the navigation controller are exhibited in Sect. 4, vehicle limitations are commented on in Sect. 5, and relevant conclusions are summarized in Sect. 6.

2 System Description

2.1 Vehicle Description

The prototype used in this work is a 4:1 scale model of the BMW i8 Spyder controlled by a Raspberry Pi 4B. This vehicle is equipped with a DC motor at the right rear tire, which provides linear thrust, and a second DC motor that controls the steering angle of the steering shaft. The distance between the axes is 70 cm, and the distance between the tires is 55 cm, as pictured in Fig. 1.

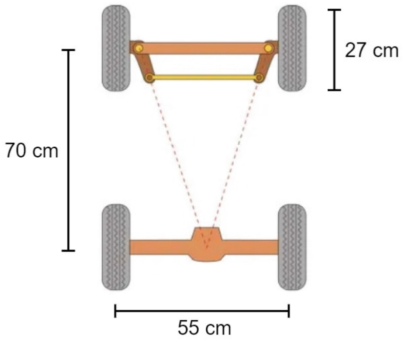


Fig. 1. Prototype dimensions.

2.2 Sensors

The vehicle is equipped with velocity and direction sensors, as well as four ultrasonic distance sensors located at the rear, front, and sides. These HC-SR04 sensors detect obstacles within a 4 m range in direct sight. The power supply consists of two 12 V 7 Ah lead batteries. The first one powers the entire electronic systems, and the second one powers the DC motors.

Linear Encoder. To measure linear velocity, a slotted wheel shown in Fig. 2a was installed at the right rear wheel. This wheel was designed to produce 90 signal transitions per rotation at the light beam from an LM393 encoder. Therefore, the linear velocity v, is measured as $v = n\pi d_w/(90T_0)$ where n is the number of encoder transitions, d_w is the wheel diameter, and T_0 is the sample time at which the transitions are measured.

Steering Position Sensor. The vehicle uses a DC motor to turn the steering column, which is equipped with a velocity reduction box. To assemble the steering position sensor, a potentiometer is used to measure the steering angle. This setup, illustrated in Fig. 2b, functions as a servomotor [13].

The signal from this potentiometer is digitized using an MCP3008 analog-to-digital converter (ADC). Since the steering angle, ψ, ranges from –30 to 30°, whereas the potentiometer signal, Pt, reads values between 0.65 and 0.35 respectively, the transduction of the voltage signal to the steering angle is performed as $\psi = -200 \cdot Pt + 100$.

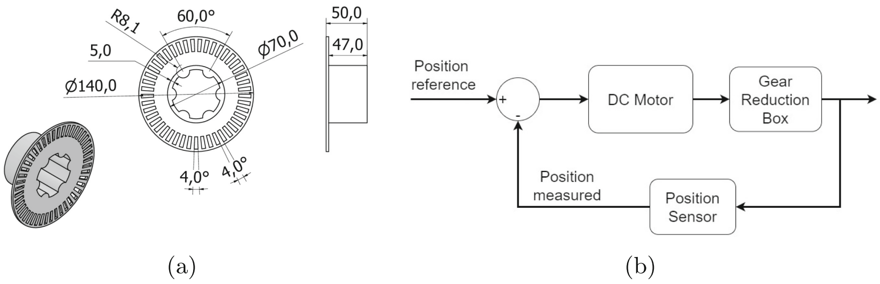


Fig. 2. Sensors installed on the system: (a) Slotted wheel blueprint. (b) Servomotor system diagram.

2.3 Architecture of the Vehicle Control System

The prototype consists of two primary subsystems: the Environment Description Subsystem (EDS) and the Navigation Subsystem (NavS). The EDS uses sensors such as LIDAR, cameras, and proximity sensors to map the environment of the vehicle. In contrast, NavS follows a predesigned trajectory, receiving data from the EDS via a CAN bus, as well as from its own navigation sensors, to control the prototype movement. Both the EDS and NavS can be remotely monitored via a WiFi connection. The general architecture of the system is shown in Fig. 3.

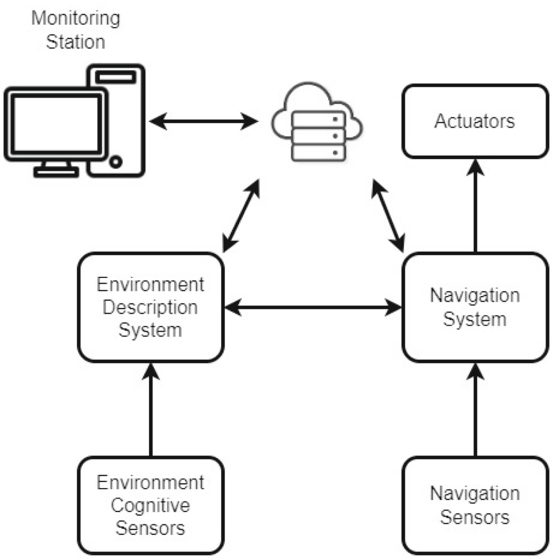


Fig. 3. System architecture diagram.

The focus of this work is the development of the main controller for the NavS. This controller comprises three embedded systems that allow the vehicle to follow a predefined path.

The NavS integrates the input of the aforementioned sensors, controls the two DC motors as actuators, and employs three embedded controllers to navigate semi-structured environments, as shown in Fig. 4. The linear algebra-based controller (LAC) processes sensor signals to calculate the optimal linear velocity and steering angle necessary to track the vehicle's target trajectory. The PID controllers then convert these commands into PWM signals that drive the DC motors. To mitigate the high error variance in raw sensor measurements, which impedes smooth vehicle control, two Kalman filters are incorporated.

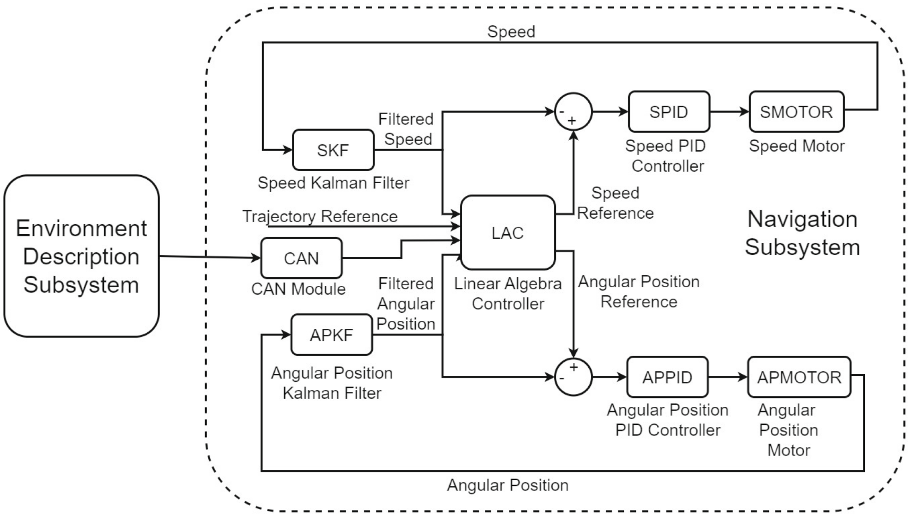


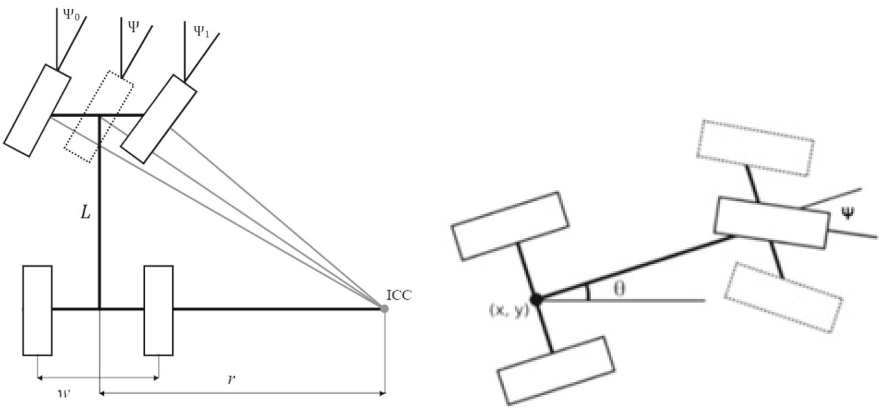
Fig. 4. Architecture of the vehicle control system.

3 Mathematical Models and System Design

3.1 Ackerman Kinematic Model

The Ackerman kinematic model, shown in Fig. 5a, describes the behavior of a four-wheeled automotive vehicle, in which the linear velocity is provided by a powertrain on the rear axis, and the steering angle is controlled by a steering column on the front axis [16]. Furthermore, by using the simpler Ackerman bicycle model [7,20], the front and rear axes can be modeled as a single wheel in the middle of the vehicle, as shown in Fig. 5b.

The vehicle position is measured in a reference coordinate system, relative to the middle of the rear axis. The vehicle's orientation, θ, is the angle at which the



(a) (b)

Fig. 5. Ackerman model: (a) Kinematic model. (b) Bicycle model.

vehicle needs to move forward to properly track the reference trajectory, and ψ is the steering angle of the front wheels [7,16].

The discrete form of the Ackerman kinematic model is defined as in (1).

$$
\begin{aligned}
x_n &= x_{n-1} + T_0 \cdot V_{n-1} \cdot \cos(\theta_{n-1}) \\
y_n &= y_{n-1} + T_0 \cdot V_{n-1} \cdot \sin(\theta_{n-1}) \\
\theta_n &= \theta_{n-1} + \tfrac{T_0 \cdot V_{n-1}}{L} \cdot \tan(\psi_{n-1})
\end{aligned} \tag{1}
$$

where x_n and y_n coordinates describe the vehicle's trajectory at instant n, while V_{n-1}, θ_{n-1}, and ψ_{n-1} refer respectively to the car speed, its orientation angle, and the steering angle of the front wheels at instant $n-1$. Furthermore, L, corresponds to the length between the front and the rear axes, and T_0, to the sampling time, set to 0.1 s.

3.2 Linear Algebra Based Controller Design

For tracking the reference trajectory, the LAC requires the current linear vehicle speed and its steering angle as inputs. Initial conditions are zero for the vehicle's position and orientation. The LAC calculates the target linear velocity and steering angle, so equations in (1) are rewritten in matrix form, as shown in (2).

$$
\begin{bmatrix} \cos(\theta_n) & 0 \\ \sin(\theta_n) & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} V_n \\ \tan(\psi_n) \end{bmatrix} = \frac{1}{T_0} \cdot \begin{bmatrix} x_{n+1} - x_n \\ y_{n+1} - y_n \\ \frac{\theta_{n+1} - \theta_n}{\frac{V_n}{L}} \end{bmatrix} \tag{2}
$$

For the system to have a unique solution, the condition in (3) needs to be satisfied.

$$\begin{bmatrix} \cos(\theta_n) \\ \sin(\theta_n) \end{bmatrix} \cdot V_n = \frac{1}{T_0} \cdot \begin{bmatrix} x_{n+1} - x_n \\ y_{n+1} - y_n \end{bmatrix} \tag{3}$$

This condition forces both direction vectors to be parallel, leading to the condition described in (4).

$$\frac{\sin(\theta_n)}{\cos(\theta_n)} = \frac{y_{n+1} - y_n}{x_{n+1} - x_n} = tan(\theta_n) \tag{4}$$

To satisfy this condition, a variable is sacrificed, so that x and y coordinates can follow the reference trajectory. This variable is the orientation, so it will be called θ_{ez_n} from now on. Therefore, the desired next position of the vehicle is calculated as described in (5).

$$\begin{aligned} x_{n+1} &= x_{ref_{n+1}} - k_x \cdot (x_{ref_n} - x_n) \\ y_{n+1} &= y_{ref_{n+1}} - k_y \cdot (y_{ref_n} - y_n) \\ \theta_{n+1} &= \theta_{ez_{n+1}} - k_\theta \cdot (\theta_{ez_n} - \theta_n) \end{aligned} \tag{5}$$

where k_x, k_y and k_θ are constants between 0 and 1 that determine the roughness of the controller action. The closer to 1, the smoother the controller action.

The relations in (6) are defined to simplify the equations of the controller.

$$\begin{aligned} \Delta x &= x_{ref_{n+1}} - k_x \cdot (x_{ref_n} - x_n) - x_n \\ \Delta y &= y_{ref_{n+1}} - k_y \cdot (y_{ref_n} - y_n) - y_n \\ \Delta\theta &= \theta_{ez_{n+1}} - k_\theta \cdot (\theta_{ez_n} - \theta_n) - \theta_n \end{aligned} \tag{6}$$

Finally, the equation system is solved by applying the linear algebra property $A^T \cdot A \cdot x = A^T \cdot b$. The solution of the system is described by (7).

$$\begin{bmatrix} V_n \\ \tan(\psi_n) \end{bmatrix} = \frac{1}{T_0} \begin{bmatrix} \Delta x \cdot \cos(\theta_{ez_n}) + \Delta y \cdot \sin(\theta_{ez_n}) \\ \frac{\Delta\theta}{V_n/L} \end{bmatrix} \tag{7}$$

3.3 Kalman Filter

The Kalman filter is an important iterative algorithm that uses predictions of the system state to reduce the error variance from measurements taken from sensors [3,21]. In this work, this filter is used to reduce the variance of both the linear speed and the steering angle measurements. The Kalman filter alternates prediction and correction (filtering) stages with each new provided measurement. It consists of five steps: projection of the state ahead and the error covariance ahead from system dynamics, computation of the Kalman gain using the state error covariance and the measurement error covariance, state updating with the help of the projected state and the current measurements, and finally, filtering of the error covariance from the projected covariance.

$$\begin{gathered} x \sim N\left(\mu, \sigma_x^2\right); x' \sim N\left(\mu, \sigma_{x'}^2\right); y \sim N\left(\mu, \sigma_y^2\right) \\ \mu = E\left\{x\right\} = E\left\{y\right\} \end{gathered} \tag{8}$$

Since the sampling time is small and we do not have a complete characterization of the dynamics of the system, the projections of the state ahead and the error variance ahead are simply chosen as the previous filtered values, as shown in (9), simplifying the model.

$$x'_k = x_{k-1}; \qquad \sigma_{x'_k}^2 = \sigma_{x_{k-1}}^2 \tag{9}$$

The filtered current system state is computed in (10).

$$x_n = x'_n + (K_n) \cdot (y_n - x'_n) \tag{10}$$

where K_n is the Kalman filter's gain, computed in (11).

$$K_n = \frac{\sigma_{x'_n}^2}{\sigma_{x'_n}^2 + \sigma_{y_n}^2} \tag{11}$$

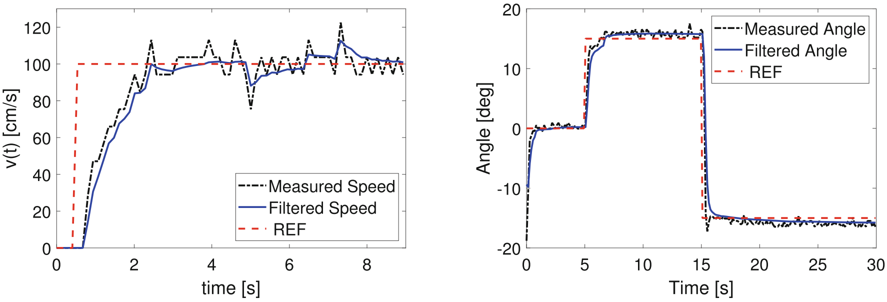
being, $\sigma_{x_n}^2$, the error covariance for the current filtered system state, shown in (12).

$$\sigma_{x_n}^2 = (1 - K_n) \cdot \sigma_{x'_n}^2 \tag{12}$$

The filtered value depends on the filter gain. As the gain is closer to one, the filtered variable is closer to the new measurement; otherwise, when the gain is closer to zero, the filtered value is closer to the projected system state ahead.

Kalman Filter for Speed Measurements. The variance of the raw speed measurements was computed, resulting in a value of 1500 $(\text{cm/s})^2$. This value was used to set up the error variance of the projected speed, while its initial value was set to 0 cm/s. The algorithm was also programmed so that when there is a significant change in the speed, the estimated variance is reset to the initial value, so that the filtered speed turns out closer to the measured value. Figure 6a shows the smoother behavior for the filtered speed in contrast to the raw data. The error variance for the system speed in the steady state was reduced from 60 $(\text{cm/s})^2$ to 30 $(\text{cm/s})^2$ once filtered.

Kalman Filter for Steering Angle Measurements. The Raspberry Pi reads the ADC output values as a fraction of the reference voltage level. When the vehicle's wheels are straight, the measured signal is 0.5, so this value is used to initialize the filter. Furthermore, the initial error variance is set to 0.1. Analogously to the speed filtering, the error variance is reset when there is a significant



(a) (b)

Fig. 6. Vehicle position variables measurement: (a) Speed measurement. (b) Steering angle measurement.

change in the steering angle to reduce the Kalman gain and achieve filtered estimates closer to the current measurements. Figure 6b shows the steering angle for raw and filtered measurements. The error variance was reduced approximately by half, from $4.38 \cdot 10^{-6}$ to $2.23 \cdot 10^{-6}$ by Kalman filtering.

3.4 Controller Area Network Communications Protocol

The Controller Area Network (CAN) protocol is a serial communication standard originally designed for the automotive industry, offering strong immunity to electrical interference and built-in error detection and correction capabilities [9]. This protocol facilitates communication between microcontrollers and electronic devices by enabling data sharing across a common bidirectional transmission bus. The CAN protocol is typically divided into layers, with the application, physical, and data layers being the most common configuration. In the standard protocol used in this application, messages consist of 10 bytes to transmit environmental information related to navigation from a secondary processor. Alternatively, the extended protocol uses 94 bits [8].

This setup is advantageous as it reduces vehicle wiring by minimizing cabled connections. The CAN bus operates as a transmission bus, meaning that all nodes connected to the bus receive transmitted data. However, message contents and arbitration rules dictate which node responds to the message [9].

In this project, a CAN bus was installed in the prototype to enable communication between the EDS and NavS. Tests demonstrated that when the NavS received a warning signal from the EDS via the CAN bus, the vehicle's trajectory shifted from the primary path to a secondary path as directed by the EDS.

3.5 Speed and Steering Angle PID Controllers

To translate the speed reference computed by the LAC, a PID controller was designed that takes as input the speed error and produces a PWM signal to feed the DC motor. The transfer function $G_v(s)$ of the speed system was identified using the Alfaro two points method [17], and it is described as in (13).

$$G_v(s) = \frac{K \cdot e^{-t_0 \cdot s}}{\tau \cdot s + 1} = \frac{3.5 \cdot e^{-0.36 \cdot s}}{2 \cdot s + 1} \tag{13}$$

where, K is the system gain, t_0 corresponds to the system delay, and τ is the system time constant. Analogously, the steering angle transfer function $G_a(s)$ is described as in (14).

$$G_a(s) = \frac{5.93 \cdot e^{-0.17 \cdot s}}{s \cdot (0.09 \cdot s + 1)} \tag{14}$$

The control action CA_{PID} provided by the PID controller is computed as in (15) [4].

$$CA_{PID} = K_C \cdot (e(t) + \frac{1}{T_i} \int e(t) \cdot dt + T_d \cdot \frac{d}{dt} e(t)) \tag{15}$$

where e(t) represents the error between the reference and the current value, and K_C, T_i and T_d are the controller's proportional, integral and derivative gain, respectively.

The speed PID controller is tuned with Dahlin's method since it produces a smooth response without an overshoot. Further, tuning expressions and the gain values for both PID controllers are reported in Table 1, where τ_c is a tuning parameter, that is chosen as in (16) [4].

$$\frac{\tau_c}{t_0} > 0.8 \text{ y } \tau_c > 0.1 \cdot \tau \tag{16}$$

Table 1. PID Controllers Tuning Methods

Gain	Speed Controller		Steering Angle Controller	
	Expression	Value	Expression	Value
K_C	$\frac{1}{2 \cdot K} \cdot \frac{\tau}{t_0}$	0.79	$\frac{2 \cdot \tau_c + \tau + t_0}{K \cdot (\tau_c + t_0)^2}$	0.5
T_i	τ	2	$2 \cdot \tau_c + \tau + t_0$	1.25
T_d	$\frac{t_0}{2}$	0.18	$\frac{(2 \cdot \tau_c + t_0) \cdot \tau}{2 \cdot \tau_c + \tau + t_0}$	0.08

To showcase the functionality of the speed PID controller, a step change from 0 to 100 cm/s is applied as a reference. Figure 6 depicts the controller's behavior. It is evident that the actual speed of the vehicle successfully tracks the reference.

Similarly, for the steering angle PID Controller, a reference with step changes to 15° and −15° are established. A slight discrepancy between the reference and the actual steering angle is exhibited, attributed to the weight of the steering column. Figure 6b depicts the behavior of the controller.

4 Results from Implementation of the LAC

A circular trajectory with a 2-m radius was designed to test the LAC system embedded with the PID controllers. The vehicle successfully followed the trajectory, achieving a root mean squared error (RMSE) of 10.1 cm, as illustrated in Fig. 7a.

Additionally, the LAC calculates the vehicle's desired orientation relative to its current position and the reference trajectory. Figure 7b shows the vehicle's orientation performance in comparison to the LAC-generated reference, with an RMSE of 0.11 rad.

The performance of the PID controllers was also evaluated during the test. The steering angle behavior, depicted in Fig. 7c, reveals that the PID controller struggles to accurately follow a variable reference, resulting in an RMSE of 19.1°.

Figure 7d shows the vehicle's speed behavior throughout the test. Speed drops were observed at certain intervals, attributed to inaccuracies in the speed sensor installed on the prototype. Nevertheless, the controller was able to compensate for this sensor error, tracking the LAC-generated reference with an RMSE of 22.94 cm/s.

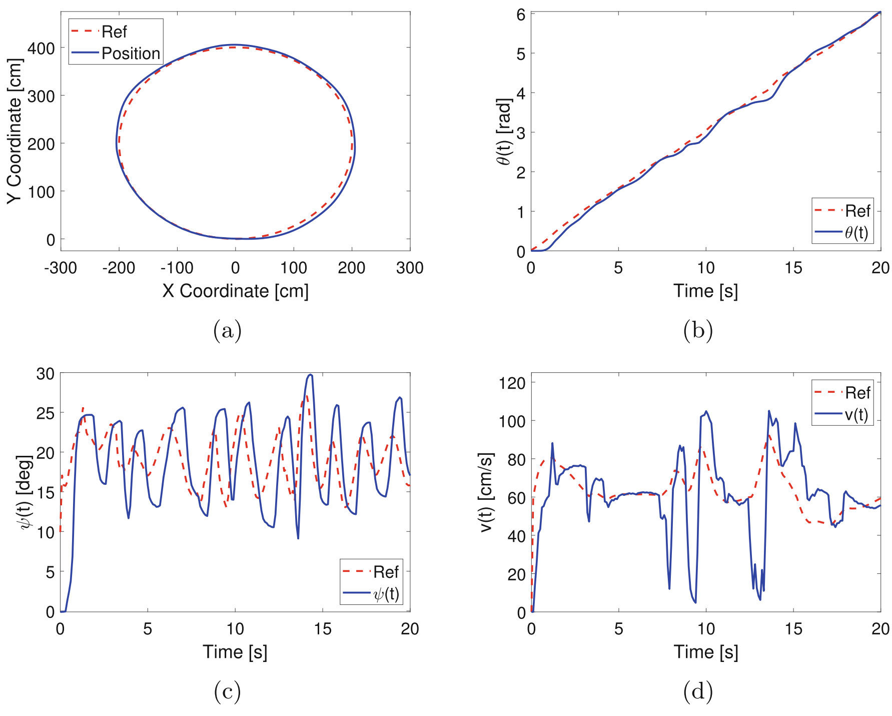


Fig. 7. Circular trajectory tracking test: (a) Trajectory tracking. (b) Vehicle orientation. (c) Steering angle. (d) Vehicle speed.

A CAN expansion module was attached to the main processor to facilitate communication between the NavS and the EDS. To demonstrate the functionality of this system, two distinct routes were programmed: a straight line and a circular trajectory with a 2-m radius, as used in the previous test. The vehicle was programmed to switch from the first route to the second upon receiving a warning message. Figure 8a illustrates the vehicle's position relative to the trajectory, showing that it successfully adjusted its path upon receiving the message, with a root mean squared error (RMSE) of 7.6 cm. The behavior of the vehicle's orientation, speed, and steering angle closely mirrors the results from the previous test.

Figure 8b presents the vehicle's orientation behavior in comparison to the LAC-generated reference. The vehicle tracks the orientation reference within an acceptable error range of 0.1 rad, with the highest error detailed in the figure.

Similarly, Fig. 8c shows the steering angle performance relative to the LAC-generated reference. The PID controller struggles to follow the variable trajectory accurately, resulting in an RMSE of 17.57° for the steering angle.

The vehicle's speed during the CAN message reception test was also analyzed. The vehicle tracks the speed reference generated by the control system, though a speed peak is observed at second five in Fig. 8d. This anomaly occurs because the test was designed to stop the vehicle completely before switching to the second trajectory, causing a brief spike upon restarting.

Additionally, Fig. 8d shows intervals where the speed drops to 0 cm/s, an error attributed to the inaccuracy of the installed speed sensor. Despite this, the PID controller for the speed subsystem compensates for the sensor error, following the generated trajectory with an RMSE of 22.9 cm/s.

5 Vehicle Limitations

The system has inherent physical and mechanical limitations that restrict its ability to navigate routes beyond its capabilities. Initially, any programmed route must be designed with a minimum traveling speed of 50 cm/s to ensure that the vehicle's motor can effectively propel its weight.

Moreover, due to the vehicle's substantial mass, increasing velocity results in higher system inertia, causing the vehicle to widen its trajectory's curvature radius. Consequently, the vehicle may not be able to return to its starting point on circular routes. Extensive trial and error determined that the vehicle can successfully complete a circular route with a minimum radius of 2 m at 50 cm/s. Furthermore, the steering angle system is limited by mechanical stoppers that prevent angles from exceeding 30°, as surpassing this threshold would strain the gear reduction box.

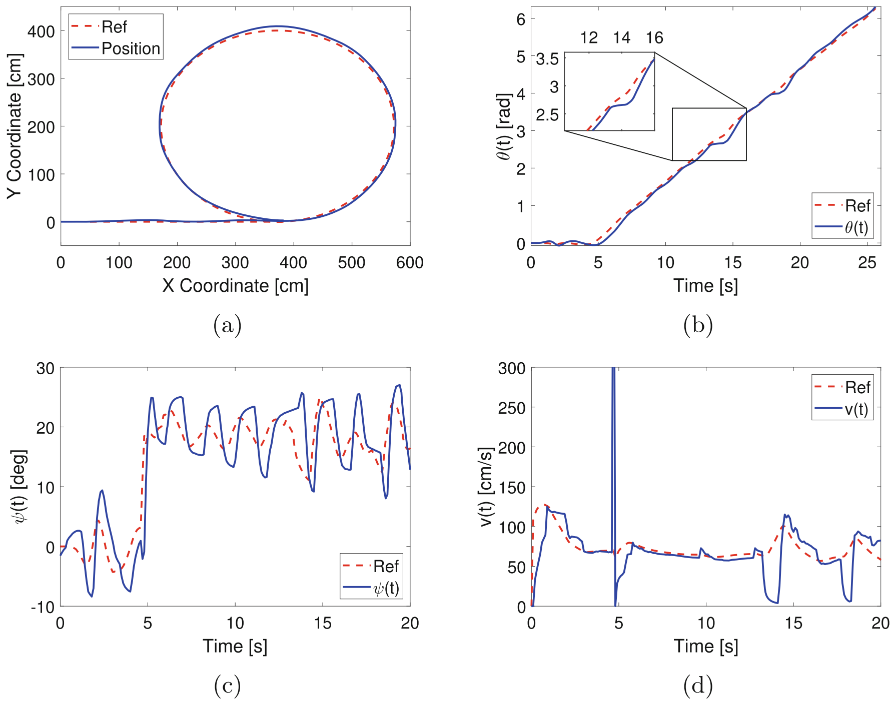


Fig. 8. CAN message reception test results: (a) Trajectory tracking. (b) Vehicle orientation. (c) Steering angle. (d) Vehicle speed.

6 Conclusions

This work presents the design of an autonomous vehicle based on a scaled model of the BMW i8 Spyder, with a wheelbase of 70 cm and a width of 55 cm. The vehicle is equipped with ultrasonic distance sensors, a velocity sensor, and a steering angle position sensor. It is powered by two DC motors, which control both linear velocity and steering column movement. Additionally, a CAN expansion module is installed on the main processor to enable communication between the Navigation Subsystem (NavS) and the Environment Description Subsystem (EDS).

The vehicle employs a linear algebra-based controller, complemented by two PID controllers, to track pre-programmed trajectories within a referential coordinate system. However, deviations between the reference and actual positions arise due to nonlinearities in the steering mechanism and the weight of the prototype.

The testing phase involved the vehicle navigating circular trajectories, demonstrating the effectiveness of the designed controllers in trajectory tracking. Additionally, a message reception test showcased the vehicle's ability to alter its trajectory upon receiving a warning signal.

While the results are promising, route design must consider the prototype's physical limitations to ensure successful navigation. The current findings provide sufficient evidence that the prototype is ready for further testing in semi-structured environments.

However, the tests reveal that the PID controller for the steering angle subsystem struggles to accurately follow the trajectory generated by the linear algebra-based controller (LAC), with the steering angle RMSE reaching nearly 20°. The authors recommend designing a control law better suited for tracking variable trajectories, such as an additional linear algebra-based controller, which has been shown to improve reference tracking in similar studies [5,18].

Despite these limitations, both tests show consistent behavior across RMSE values for trajectory tracking, vehicle orientation, steering angle, and speed.

Acknowledgments. The authors thank Dr. Oscar Camacho and CEDIA for facilitating international cooperation to complete this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azizi, M.R., Rastegarpanah, A., Stolkin, R.: Motion planning and control of an omnidirectional mobile robot in dynamic environments. Robotics **10** (2021). https://doi.org/10.3390/robotics10010048
2. Barbosa, J., Hernández, C., Paredes, D., Játiva, R., Villarreal, K.: Diseño e implementación de un prototipo de vehículo con capacidades de navegación autónoma y de exploración cognitiva de su entorno. Desarrollo e Innovación en Ingeniería **2021**(1), 233–252 (2021)
3. Becker, A.: Kalman Filter From The Ground Up, 2nd edn (2023)
4. Camacho, O., Rosales, A., Rivas, F.: Control de Procesos. Escuela Politécnica Nacional, 1 edn. (2020)
5. Camacho, O., Scaglia, G., Lucia Quintero, O.: A dead time compensator based on linear algebra (dtcla). IFAC-PapersOnLine **50**(1), 3075–3080 (2017)
6. Cheein, F.A., Scaglia, G.: Trajectory tracking controller design for unmanned vehicles: a new methodology. J. Field Rob. **31**, 861–887 (2014). https://doi.org/10.1002/rob.21492
7. Conner, D.C., Kress-Gazit, H., Choset, H., Rizzi, A.A., Pappas, G.J.: Valet parking without a valet. In: IEEE International Conference on Intelligent Robots and Systems, pp. 572–577 (2007). https://doi.org/10.1109/IROS.2007.4399374
8. Cook, J.A., Freudenberg, J.S.: Controller area network (can) (2008)
9. Corrigan, S.: Introduction to the controller area network (can) application report introduction to the controller area network (can) (2002)
10. Faisal, A., Yigitcanlar, T., Kamruzzaman, M., Currie, G.: Understanding autonomous vehicles: a systematic literature review on capability, impact, planning and policy. J. Transp. Land Use **12**, 45–72 (2019). https://doi.org/10.5198/jtlu.2019.1405
11. International, S.: SAE levels of driving automation™ refined for clarity and international audience (2021). https://www.sae.org/blog/sae-j3016-update

12. Jiang, C., et al.: R2-rrt*: reliability-based robust mission planning of off-road autonomous ground vehicle under uncertain terrain environment. IEEE Trans. Autom. Sci. Eng. **19**(2), 1030–1046 (2022). https://doi.org/10.1109/TASE.2021.3050762
13. Kinch, R.: Converting an ordinary dc motor to a servomotor. https://www.truetex.com/servomod.htm
14. Li, D., Jiang, Q., Li, S., Liu, X., Liu, J.: Networked edge intelligence for autonomous farm vehicles. IEEE Netw. **37**(4), 297–304 (2023). https://doi.org/10.1109/MNET.006.2300021
15. Medina, L., Guerra, G., Herrera, M., Guevara, L., Camacho, O.: Sliding mode control approaches applied to trajectory tracking for non-holonomic mobile robots. Int. J. Control Autom. Syst., 1–17 (2022). https://doi.org/10.1007/s12555-xxx-xxxx-x
16. Patnaik, A., et al.: Design and implementation of path trackers for Ackermann drive based vehicles. arXiv preprint arXiv:2012.02978 (2020). https://doi.org/10.48550/arXiv.2012.02978
17. Ruíz, V.M.A.: IdentificaciÓn de procesos sobreamortiguados utilizando tÉcnicas de lazo cerrado. Revista Ingeniería **11** (2011). https://doi.org/10.15517/ring.v11i1-2.604
18. Sardella, M.F., Serrano, E., Camacho, O., Scaglia, G.: Linear algebra controller design based on the integral of desired closed-loop behavior: application to regulation and trajectory tracking in a typical chemical process. Ind. Eng. Chem. Res. **59**(45), 20131–20140 (2020)
19. Schwarting, W., Alonso-Mora, J., Rus, D.: Planning and decision-making for autonomous vehicles. Ann. Rev. Control Robot. Auton. Syst. **2018**(1), 187–210 (2018). https://doi.org/10.1146/annurev-control-060117
20. Sotelo, M.A.: Lateral control strategy for autonomous steering of Ackerman-like vehicles. Rob. Auton. Syst. **45**, 223–233 (2003). https://doi.org/10.1016/j.robot.2003.09.002
21. Ulin-Avila, E., Ponce-Hernandez, J.: Kalman filter estimation and its implementation (2021)
22. Yamakawa, S., Ebara, K.: Control of mobile robot by switching traveling direction and control gain. ROBOMECH J. **4**(1), 1–10 (2017). https://doi.org/10.1186/s40648-017-0097-z

Using Machine and Deep Transfer Learning for Classification of EEG Signals from Embodied and Non-embodied Priming in a Motor Imagery Training in Virtual Reality

Michael S. Ramírez-Campos[1](✉), Hamed Tadayyoni[2], and Alvaro D. Orjuela-Cañón[1]

[1] School of Medicine and Health Sciences, Universidad del Rosario, Bogota D.C., Colombia
{michaels.ramirez,alvaro.orjuela}@urosario.edu.co
[2] Faculty of Health Sciences, Ontario Tech University, Oshawa, ON L1G 0C5, Canada

Abstract. Motor imagery (MI) frameworks have a long history of being used in different motor training applications. Since the emergence of brain-computer interfaces (BCI), MI techniques have been integrated into BCI frameworks to enable controlling external devices through interpreting neural signals into executable commands. Virtual reality (VR) has the capacity to introduce novel ways of improving the performance of MI-BCI trainings through enabling embodiment prior to and during the tasks. While the performance improvements achieved by using VR-based embodiment during the MI training has been investigated previously, the effects of VR-based motor priming prior to the training needs to be further addressed. This study uses machine learning (ML) to find out whether or not VR-induced avatar embodiment before the actual MI-BCI training is capable to make significant differences in electroencephalography (EEG) signals recorded from the users. Detecting the most relevant features which best represent such differences enables the introduction of biomarkers of VR-based motor priming embodiment in MI-BCI applications. Relating these biomarkers to the specific neurophysiological functions which facilitate MI can help in design and development of MI-BCI applications with improved accuracy and performance.

Keywords: brain-computer interface · motor imagery · embodiment · virtual reality · electroencephalography · machine learning

1 Introduction

Defined as the mental representation of a movement action without actually performing it [1], motor imagery (MI) techniques have been used in various domains of motor training to enhance training and performance. This is achieved through stimulating the related neural pathways via mental practice and visualization of the intended movements. The MI trainings result in strengthening the neural pathways in the brain associated with motor skills, enhancing motor accuracy and improving the overall motor performances

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 140–153, 2025.
https://doi.org/10.1007/978-3-031-88854-0_10

related to the discipline where the training is being used [1–5]. The applications of MI include (but are not limited to) rehabilitations, sports, cognitive, different types of motor therapies, and even music [2]. For example, in sports psychology, MI – alongside other mental practices- has been proved to be affective on the acceleration of learning and improvement of motor skills, increasing confidence in the athletes and improving their techniques [3]. In rehabilitation, MI is utilized as a powerful tool to facilitate relearning or restoring motor skills after a stroke, injury or illness [4–6]. This is mainly due to the fact that the event-related synchronization/desynchronization (ERS/ERD) during an MI task resembles the activities of the brain during the actual planning and practice of the movements. As a result, MI activities helps in strengthening and restoration of the neural pathways related to motor functions [6, 7].

A more recent application of motor imagery is in brain-computer interfaces (BCI) where it enables individuals to control external devices through an interface that is capable of interpreting the MI-associated neural signals [8]. In such applications, motor imagery is generated by the users through envisioning certain movements (such as moving fingers, hands, etc.) while not doing the actual physical movements. The related neural activities of the users' brain are then detected through some type of neuroimaging (e.g. electroencephalography (EEG) [9], functional magnetic resonance imaging (fMRI) [10], etc.) and then interpreted into predefined commands for controlling a device through its interface. This not only offers creative solutions to people with motor conditions and disabilities, but also introduces new ways of communicating with motor interfaces for the healthy population.

Effective detection of MI cues and their translation into BCI commands requires recognition of distinct patterns in the signals related to certain motor activities/commands depending on the utilized neuroimaging technique. For example, when using EEG signals for an MI-BCI interface, it is important to consider how variable certain characteristics of the signals used in BCI applications may be for the same motor imagery task among different individuals [11, 12]. This requires extensive training frameworks for the users to become familiar with the process and become capable of generating more robust EEG patterns based on the visual and performance-based neurofeedback that they receive from the interface [13]. All these drawbacks have resulted in introduction of new paradigms into MI-BCI, including the application of virtual reality (VR).

VR has the capacity to offer a more real-time and immersive feedback to the MI-BCI user, enabling an immediate sense of performance in terms of the accomplished motor task [14]. Providing a type of visual feedback that resembles the natural setting of performing the intended motor task, VR allows the users to see whether the intended task is done correctly while performing it, resulting in higher performance rates than other visual feedbacks (such as flat screen feedbacks) for MI-BCI tasks [15, 16].

Moreover, the performance of MI-BCI trainings can be enhanced through motor priming-induced learning. Motor priming is the process of offering stimulant cues for activating specific neural pathways related to that motor task in the user's brain [17]. Motor priming results in increasing MI performance through facilitating the neural plasticity in the user's brain as well as increasing their motivation, attention and engagement in the intended task [17, 18]. Motor priming can be introduced to the learning process through different sensory cues, such as visual, auditory and somatosensory stimuli.

Auditory priming involves playing sudatory instructions as well as sound effects related to the motor task, helping the users to synchronize their imagined movements with the presented auditory stimuli [17, 19]. On the other hand, somatosensory priming includes introducing proprioceptive and/or tactile stimuli and feedback to the MI task, resulting in the users to better imagine the MI task and enhance their performance [20]. One of the most effective types of priming is visual priming which involves introducing relevant visual stimuli to the MI training. This can be achieved through showing the footage of related body movements before asking the users to start their MI task, activating their visual cortex.

With the capability to present visual, auditory, tactile and proprioceptive feedbacks, VR offers a unique opportunity to combine different types of priming-induced learning to the MI-BCI frameworks. Additionally, it can introduce physical activity related to the MI task which can further boost the performance of the training [21]. All these characteristics enable VR to induce a unique sense of embodiment to the MI-BCI prior to and during the training. The VR-induced sense of embodiment might be helpful to the BCI performance [22] and how the resulting MI-related EEGs are modulated by the BCI framework [23].

While there has been an open debate about the effectivity of sense of embodiment embedded in the MI-BCI training on the BCI performance and the modulation of the MI-related EEG characteristics, its impact when it is used as a motor priming phase prior to the actual MI-BCI task has never been studied before [9]. The goal of that study was to investigate the effects of the VR-based sense of embodiment induced during the priming phase prior to the actual MI-BCI training on the enhancement of the ERD patterns. This between-subjects study concluded that there is no significant difference between the ERD patterns and MI-BCI accuracy between the experiment group who were subjected to a VR-based embodiment motor priming phase and the control group who were deprived of such embodiment phase prior to the beginning of the MI-BCI task. While that conclusion was made based on training a linear classifier using a standard BCI feature extraction method emphasizing on comparison of lateralization index, investigating the differences of the EEG features using other machine learning (ML) algorithms and through classification based on other EEG-based features is not investigated.

In this study, we leverage a dataset collected in [9] to find biomarkers of VR-induced sense of embodiment prior to BCI training in the EEG signals recorded during the MI-BCI task [24]. Several different ML algorithms (Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC), Decision Tree (DT), Naive Bayes (GNB), K-Nearest Neighbors (KNN), Random Forest (RF) and a Multilayer Perceptron (MLP)) were used to classify the EEG signals recorded during the MI-BCI task for the two experiment and control groups mentioned before. The study aimed to determine whether ML approaches were able to accurately classify the two groups based on the features extracted from EEG data, and determine which features best represent the effects of priming sense of embodiment on the MI-BCI.

Additionally, we explore the potential of using neural networks for classifying the two groups based on the scalograms generated of the EEG signals for different time windows through Continuous Wavelet Transform. Contrary to the traditional machine learning methods, where various time and frequency domain features are offered to be used to train the model and assess the accuracy of the resulting classifiers, in neural networks, we use scalograms as raw data to be fed into the algorithms without the need for manual feature engineering [25]. Therefore, in this study, we incorporate deep (EfficientNet, MobileNet V2 and NasNet) and shallow neural networks to see how they can automatically discover complex hidden patterns and automatically learn intricate features from the scalograms. Although the usage of the mentioned pre-trained neural networks in EEG classification through scalograms may be limited due to some reasons such as the need for fine-tuning of the networks for the specific application, the results in this study are promising due to the ability of these architectures to automatically extract complex patterns and hierarchical features that shows potential in capturing meaningful information from time-frequency representations.

If accurately classified, the features that mark the difference of the EEG signals among the two groups can be used as biomarkers of priming sense of embodiment during the MI-BCI tasks. Finding and studying the neurophysiological origins of these markers helps better understand the effects of sense of embodiment on the performance of VR-based motor imagery tasks and also will be helpful in developing novel methods for detection and interpretation of MI-induced neural activities in the brain and designing better BCI frameworks.

2 Methods

2.1 Dataset Demographics and Experimental Design Characteristics

In this study, we take advantage of using the dataset (EEG and EMG) collected by Vagaja et al in [9]. The dataset comprises of EEG and EMG signals of 26 participants consisting of 10 male (mean age 25.4 ± 7.4 years old) and 16 females (mean age 23 ± 3.2 years old), all right-handed with normal vision and no reported motor impairments. Half of the participants were randomly assigned to either the experiment or control groups. The signals were recorded during two phases of the study: 5-min priming and 10-min motor imagery training. The priming phase involved two different setups between the two experiment and control groups. The experiment group underwent an embodiment priming in which, for the first 3 min, they were instructed to move their upper limbs, head and upper body and their first-person VR avatar matched their movements in synchrony. For the subsequent 2 min, a soft paint brush was used to induce the virtual hand illusion in them through stroking their real right hand in synchrony with a similar animation for the right-hand of their virtual avatar. On other hand, the control group's first 3 min of priming included them watching a footage of a pre-recorded third-person avatar in VR that did not move in synchrony with their movements and their subsequent 2 min involved their left hand being stroked by the brush while the third-person avatar's right hand was stroked.

The MI-BCI training phase involved all participants imagining the sensation of making a fist in their hand based on an arrow pointing to either their first-person VR avatar's left or right hand. This training consisted of 40 10-s trials (20 trials distributed randomly between the left and right hands) followed by a 10-s resting period.

The EEG signals were recorded through a wireless EEG amplifier (sampling rate of 500 Hz) and the EEG cap consisted of 32 active electrodes (+3 ACC) distributed spatially by the 10–20 EEG standard.

2.2 Signal Pre-processing

The whole signals were filtered using a bandpass filter of 0.1 Hz and 60 Hz with a 24 dB/octave steepness slope, and a notch filter of 50 Hz for removing the noise associated to electrical noise. This filtering approach allowed us to remove the eyeblinks given their nature of low frequency.

2.3 Machine Learning Pipeline

To investigate the effects of temporal window length on the accuracy of each ML model, the EEG signals were segmented into 1, 2, 5, and 10 s windows. Then all segmented windows were labelled as experimental (embodied) or control (non-embodied). Later, for each segmentation length, time-domain, frequency-domain and non-linear features were extracted from the segmented EEG signals groups. For feature selection, maximum relevance minimum redundancy (MRMR) combined with Mann-Whitney U test was used [26].

The features employed in this study were predominantly chosen by drawing upon previous research findings that demonstrated efficacy in identifying optimal features for machine learning (ML) classification of electroencephalogram (EEG) data across various applications [27, 28]. The time-domain features encompass Temporal Variance (activity), Mobility, and Complexity, while the frequency-domain feature comprises power spectral density. Additionally, other features incorporated in the analysis consist of Entropy, Permutation, Spectral Entropy, as well as nonlinear measures such as Higuchi's Fractal Dimension and Hurst's Exponent. Statistical characteristics like Kurtosis and Skewness were also included as part of the feature set. The combination of EEG channels of interest (FC1, FC2, FC5, Cz, C4, CP1, CP2, CP5, CP6), the basic features (10), and the frequency bands (delta, theta, alpha, beta and gamma) made the final count of candidate features to be $9 \times 10 \times 5 = 450$.

The study employed various classification techniques including Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC), Decision Tree (DT), Naive Bayes (GNB), K-Nearest Neighbors (KNN), Random Forest (RF), and Multilayer Perceptron (MLP). Subsequently, a heuristic approach was employed to optimize the training hyperparameters of these models. The dataset was divided into 80% for training and the remaining portion for testing. Post-classification, the study identified and introduced relevant markers consisting of channel-frequency band combinations that offer significant information through specific features to determine whether the signal belongs to the recordings from the embodied or non-embodied data groups. The performance of the classifiers was evaluated based on the accuracy percentage metric.

Moreover, to incorporate the neural networks for this type of data, for each trial in the training and for each participant (in the experiment and control groups), a certain number of scalograms should have been generated. To maintain consistency with the traditional machine learning classifications used in this study, the EEG signals were

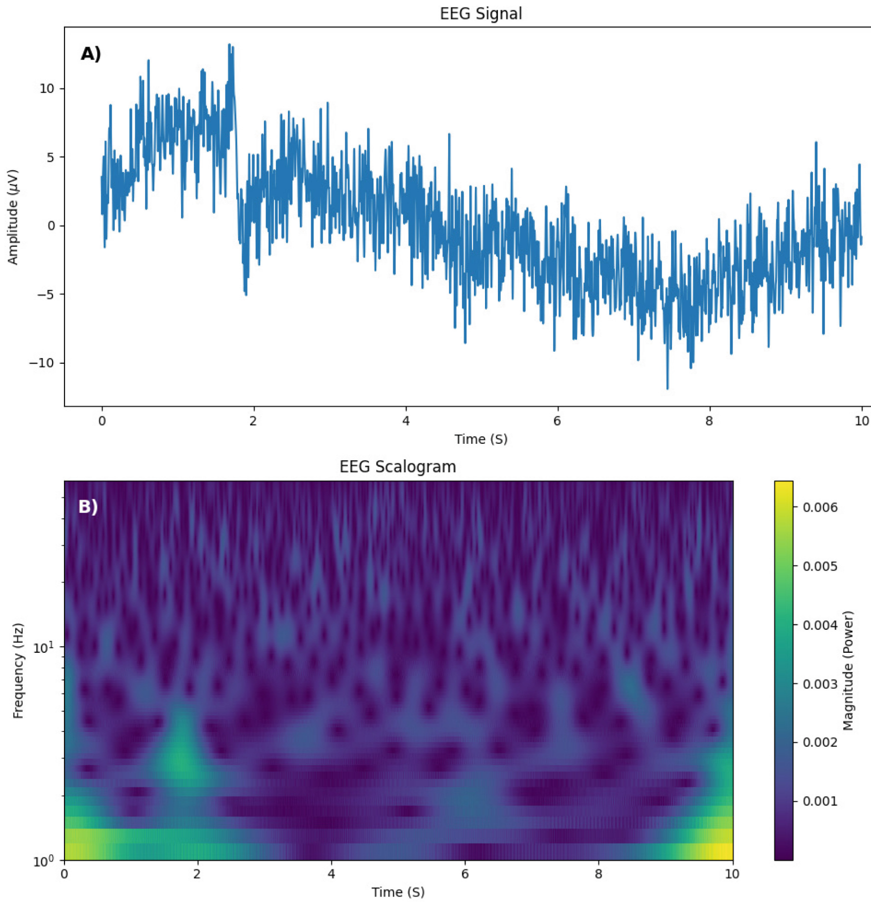


Fig. 1. A) EEG signal extracted during one of the 10-s tests performed by the participants. B) Scalogram corresponding to the EEG signal shown in figure A.

segmented into 2, 5 and 10 s windows and for each one, a scalogram was generated through Continuous Wavelet Transform. In this method, a scalogram represents how signal's power is distributed among different frequency components at each moment in time. It is worth mentioning that wavelet analysis is especially suitable for non-stationary signals such as EEG, as it provides a detailed time-frequency representation. In contrast to the Fourier Transform, which delivers frequency content averaged across the entire signal, the Wavelet Transform captures localized frequency variations over time, making it a valuable tool for identifying dynamic changes in EEG activity. Subsequently, the time-frequency variations in the generated scalograms allows the utilized neural network to discover complex patterns that can be used to classify the EEG signals. An example of a 10 s EEG time window and its corresponding scalogram is shown in Fig. 1.

3 Results

Figures 1 and 2 show the performance of the classifiers for each temporal window length group (10, 5, 2 and 1 s in length) in terms of percentage accuracy. Two different batches were used for training and testing the classifiers. All 450 features and the top 10% most relevant features (which are the first 45 features) were selected using the MRMR complemented by using the Mann-Whitney U test.

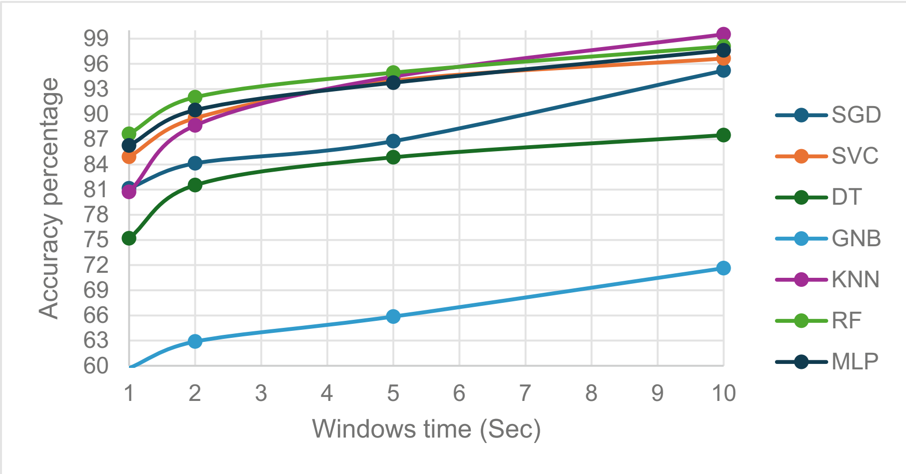


Fig. 2. Classifiers performance for all features (450) set.

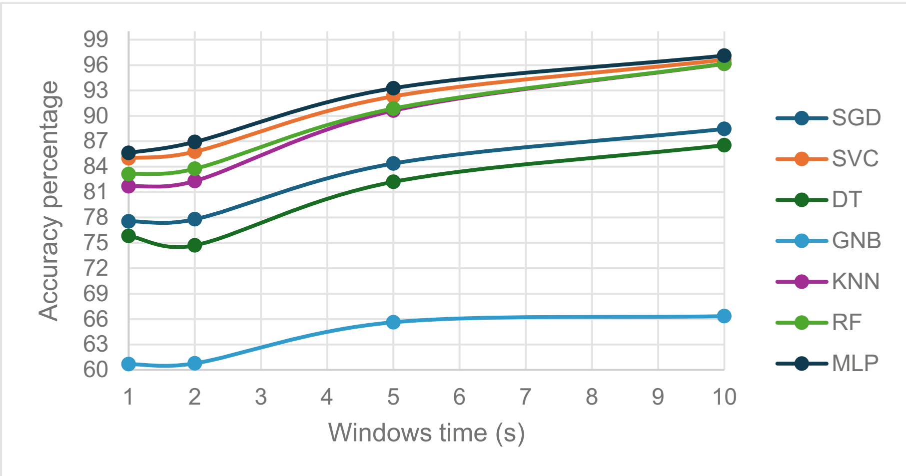


Fig. 3. Classifiers performance for the 45 most relevant features (10% of the total) set.

Since the best performances were presented when using the features extracted from the 10 s windows group, it was decided to perform additional classifications in order to test the most relevant features and thus propose them as possible biomarkers. Table 1 shows the percentages of accuracy obtained by each classifier for different sets of relevant features extracted from the 10 s swindows group.

Table 1. Classifiers performance for different number of best relevant features extracted from 10-second windows

	5 features	10 features	20 features	45 features	All features
SGD	70,67	81,25	83,17	88,46	95,19
SVC	72,12	85,58	90,87	96,63	96,63
DT	73,56	80,77	80,77	86,54	87,50
GNB	63,46	64,42	76,92	66,35	71,63
KNN	72,12	87,50	90,87	96,15	99,52
RF	71,63	85,58	88,94	96,15	98,08
MLP	69,71	87,50	91,83	97,12	97,60

Table 2 presents the *p* value of 10 most relevant features extracted from the 10 s windows group resulting after applying MRMR + Mann Whitney. Figure 3 demonstrates the difference between the mean values of the best features in the most relevant frequency bands for each class. As shown in Fig. 3 the activity in theta had CP5, CP2 and CP1 as relevant channels. For the PSD in theta CP5, CP2 and CP6 are the relevant channels. To make it easier to visualize the differences in the figure, all the features' values were normalized between 0 and 1, and each pair of topographic maps has its own color scale.

For the neural networks used in this study, the number of scalograms used for each type of temporal window length in each group = the number of participants (in each

Table 2. P-value for the top 10 most relevant features

Feature	P value
CP5_4-8Hz_activity	2,21E-30
CP5_4-8Hz_psd	4,36E-30
CP2_4-8Hz_psd	1,77E-26
CP2_4-8Hz_activity	2,18E-25
FC5_8-12Hz_psd	3,41E-24
FC5_8-12Hz_activity	5,95E-24
Cz_30-60Hz_activity	2,87E-22
CP6_4-8Hz_psd	4,11E-22
FC2_30-60Hz_psd	8,96E-22
CP1_4-8Hz_activity	8,96E-22

group) × 40 trials × (10 s duration of each trial/window length) × 9 channels of interest. To address the high computational complexity and resource demands associated with running the neural networks for a huge number of scalogram images, in this part, the 1-s window group were excluded from the neural network analysis and the RGB scalogram images were converted to grayscale images. Table 3 summarizes the accuracy percentages for the neural networks for the scalogram images. Additionally, the optimizers used for training different neural network architectures when classifying EEG-based scalograms which were found through heuristic search in hyper parameters are written in Table A.1 in the appendix (Fig. 4).

Table 3. Classifiers performance for different neural networks corresponding to different windows lengths

Window Length	Shallow CNN	EfficientNet	MobileNet V2	NasNet
10 s	85,12	83,65	80,29	75,96
5 s	83,41	82,45	66,58	66,66
2 s	77,88	69,62	68,65	64,12

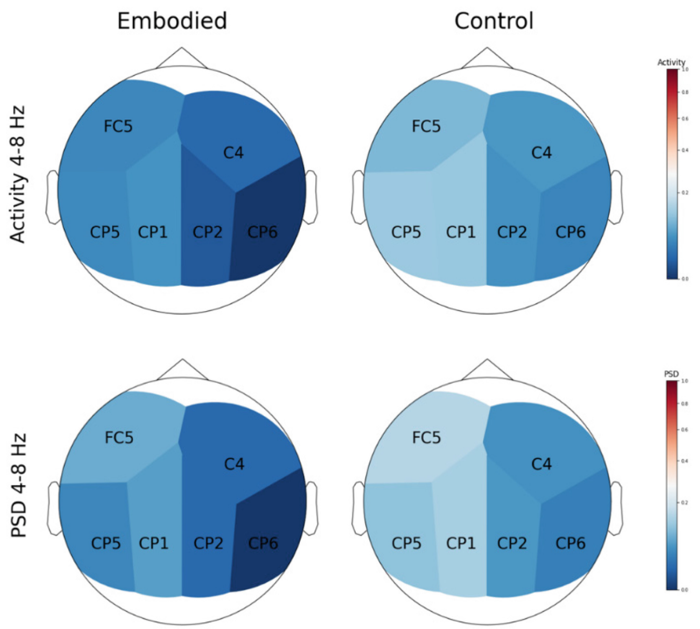


Fig. 4. Topographic map for the most relevant features and frequency bands.

4 Discussion

Since the best performances were presented when using the features extracted from the 10 second windows group, it was decided to perform additional classifications in order to test the most relevant features and thus propose them as possible biomarkers. Table 1 shows the percentages of accuracy obtained by each classifier for different sets of relevant features extracted from the 10 second windows group.

This approach included a total of 450 features and differentiated the experimental group from the control group obtaining accuracy percentages for each of the window groups above 88% when using the set of all features (Fig. 1) and 85% when using the most relevant features representing 10% of the total data (Fig. 2). The 10-s window group yielded the best results through KNN classification with an accuracy rate of 87.5% using the top 10 most relevant features and going up to 99.52% when using the total features (Table 1). The top 10 most relevant features include activity and PSD in Theta, Alpha and Gamma, for channels CP5, CP2, FC5, Cz, CP6, FC2, CP1 (Table 2 and Fig. 3). Figure 3 shows that the mean values of PSD and activity of the most relevant features were higher for the control group, i.e. the control group showed lower activity and lower power relative to the experimental group.

Figures 1 and 2 depict that the highest accuracy percentage is achievable through the classifiers that use the features extracted from the 10-s segmented windows. Except for the 1-s window classifiers, the percentage of accuracy increases almost linearly when the length of the segmented window increases.

This would indicate: Firstly, for this specific application, testing with a longer window length should be considered in order to find the optimal length. Secondly, using shorter window lengths (1 or 2 s) may result in the loss of important information from the signal. Thirdly, this is a novel contribution since, to the best of our knowledge, this is the first time that the length of windows is being studied to find the optimum length for having the most accurate model. As in the original work [9], the trials were 10 s long, with the current experimental design, it is not possible to investigate the performance of classifiers using the features extracted from larger segmented windows, however, for the specific exercise carried out it, 10 s window length turned out to be the optimal one within the available possibilities.

Based on our results, the most relevant features included in this study proved to be useful, since with only 10% of the total number of features we were able to obtain an accuracy rate of 97.12%. Additionally, with 20 features (almost 5%) and with 10 (almost 2.5%) it was possible to obtain 91.83% and 87.5% respectively. In this sense, we recommend using these features since in addition to being functional for the desired classification, they can provide valuable information and reduce the computational cost related to the processing of large amounts of features. The above also demonstrates the great usefulness of the feature selection method used in the present study.

When all features were used as inputs, out of the seven classifiers, five had accuracy percentages above 80%, and when the 45 most relevant features were used, four out of seven were above that level of accuracy. This represents a novel contribution in terms of the models produced, not only because with a reduced number of features an accurate classification can be made between the experimental and control groups, but also because the obtained results exceed those reported in related works. According to

Table 2 and Fig. 3, the top ten features that enable the classification are introduced as possible biomarkers of VR-based priming sense of embodiment in MI-BCI training. Figure 3 shows the neurophysiological origins of some of the most notable features.

For the neural networks analysis, the classification results reveal interesting patterns regarding the performance of different models across varying windows lengths. The shallow network surprisingly outperformed the more complex architectures across all windows lengths, with its highest accuracy of 85.12% achieved with 10-s windows. EfficientNet followed closely, showing strong performance with 10-second and 5-s windows (83.65% and 82.45%, respectively), but its accuracy significantly dropped when applied to 2-s windows. MobileNet V2 and NasNet, despite their advanced architectures, showed lower performance overall, particularly at shorter windows lengths. These results suggest that while deeper networks like NasNet and MobileNet V2 have the potential for high accuracy in other domains, they may not be as well-suited for shorter time segments of EEG-based scalograms without further fine-tuning. In contrast, the shallow network's simplicity appears to allow it to better adapt to the temporal nature of the EEG data, especially with longer time windows. This highlights the importance of both windows lengths and model architecture in achieving optimal classification accuracy for EEG data.

The comparison between the neural network-based results and the traditional machine learning methods shows that traditional machine learning methods, such as SVC, KNN, and Random Forest, significantly outperformed the pre-trained neural networks, particularly when using larger feature sets. Models like KNN and MLP achieved accuracies as high as 99.52% and 97.60% with all features, while the best neural network performance, 85.12% by the shallow network with 10-second windows, lagged behind. Traditional methods demonstrated superior classification accuracy, especially when leveraging feature selection techniques, while deeper neural networks like NasNet and MobileNet V2 struggled, particularly with shorter windows lengths. However, fine-tuning neural networks could enhance their performance, and their ability to automatically learn features without manual feature extraction remains promising for future work. This, combined with the use of pre-trained networks, eliminates the need for manual feature selection, making neural networks a valuable approach as they are further optimized.

Using pre-trained networks based on Image-Net dataset for classifying EEG signals is limited in the literature. While this makes the comparison between this study and the existing counterparts difficult, looking at the handful examples shows that the resulting accuracies in this study are resembling those of the similar works, emphasizing the potential of using neural networks and transfer learning in EEG classification for MI-BCI applications. As an example, the maximum accuracy achieved in [29] using SVM-enhanced EfficientNet for classification of grayscale scalograms is 70%. In [30], using MobileNet achieved a maximum accuracy of 79.45% for classification of scalograms of EEG signals resulted by using Continuous Wavelet Transform. The ability to achieve comparable, if not superior, performance through transfer learning highlights the promise of using pre-trained networks in motor imagery tasks, particularly for MI-BCI applications, where feature extraction from raw EEG data is challenging.

5 Conclusions

This article presents a novel and promising contribution to the research related to studying the effects of VR-based embodiments on the motor imagery brain computer interfaces. Our results show that it is possible to differentiate between two groups that performed the same task in VR and that underwent different motor priming VR-based embodiment prior to the actual training. This distinction is possible with an accuracy rate of 99.52% using a total of 450 features or with an accuracy rate of 87.5% using only the 10 most relevant features. Additionally, by investigating the effects of using different lengths for window segmentation of EEG signals to extract the features, it is found out that classifiers using the features extracted from 10-s segmented windows show the best percentage of accuracy among other classifiers. The best features which enable such classification are introduced as biomarkers of VR-based motor priming embodiment in MI-BCI applications. Relating the specific EEG frequency bands, EEG electrodes and mathematical characteristics of the EEG signals that these features represent will be helpful in further understanding the neurophysiological origins of MI-related embodiment and can be used in designing and developing MI-BCI frameworks with higher accuracy and better performance rates.

One of the novel contributions of this paper is the application of EfficientNet, MobileNet V2, NasNet, and Shallow neural networks for classifying EEG signals through scalograms using pre-trained neural networks, an approach not extensively explored in EEG classification literature. While these neural networks have demonstrated success in related fields such as neuroimaging, this study extends their application to EEG-based brain-computer interface tasks. The successful implementation of these models, despite their pre-training on non-biomedical datasets, highlights the potential for leveraging pre-trained deep learning architectures in EEG signal processing with further adaptation and fine-tuning.

The limitation of using a maximum length of 10 s for the windows should be addressed in the future works by an experimental design that allows longer EEG recordings. Furthermore, future research endeavors can explore the potential of employing deep learning methodologies as a more advanced approach compared to traditional machine learning techniques used in this work to introduce features that may represent the MI-related embodiment neural pathways better in the MI-BCI applications.

Acknowledgments. Authors acknowledge the support of the Universidad del Rosario for funding this project. In addition, the contribution of research seedhub team *Semillero en Inteligencia Artificial en Salud:* Semill-IAS.

Appendix

The following table shows the optimizers used during the training of the neural networks in this study.

Table A.1. Optimizers used in training the neural networks

Window Length	Shallow CNN	EfficientNet	MobileNet V2	NasNet
10 s	Nadam	Nadam	SGD	Adam
5 s	Nadam	Nadam	Nadam	Nadam
2 s	Adam	Adam	Adagrad	Adam

References

1. Lebon, F., Collet, C., Guillot, A.: Benefits of motor imagery training on muscle strength. J. Strength Cond. Res. **24**(6), 1680–1687 (2010)
2. Schuster, C., et al.: Best practice for motor imagery: a systematic literature review on motor imagery training elements in five different disciplines. BMC Med. **9**, 1–35 (2011)
3. Feltz, D.L., Landers, D.M.: The effects of mental practice on motor skill learning and performance - a meta-analysis. J. Sport Psychol. **5**(1), 25–57 (1983). https://doi.org/10.1123/jsp.5.1.25. (in English)
4. Stevens, J.A., Stoykov, M.E.P.: Using motor imagery in the rehabilitation of hemiparesis. Arch. Phys. Med. Rehab. **84**(7), 1090–1092 (2003). https://doi.org/10.1016/S0003-9993(03)00042-X. (in English)
5. de Vries, S., Mulder, T.: Motor imagery and stroke rehabilitation: a critical discussion. J. Rehabil. Med. **39**(1), 5–13 (2007). https://doi.org/10.2340/16501977-0020. (in English)
6. Zimmermann-Schlatter, A., Schuster, C., Puhan, M.A., Siekierka, E., Steurer, J.: Efficacy of motor imagery in post-stroke rehabilitation: a systematic review. J. Neuroeng. Rehabil. **5**, 8 (2008). https://doi.org/10.1186/1743-0003-5-8. (in English)
7. Jackson, P.L., Lafleur, A.F., Malouin, F., Richards, C., Doyon, J.: Potential role of mental practice using motor imagery in neurologic rehabilitation. Arch. Phys. Med. Rehab. **82**(8), 1133–1141 (2001). https://doi.org/10.1053/apmr.2001.24286. (in English)
8. Aggarwal, S., Chugh, N.: Signal processing techniques for motor imagery brain computer interface: a review. Array **1**, 100003 (2019)
9. Vagaja, K., Linkenkaer-Hansen, K., Vourvopoulos, A.: "Avatar embodiment prior to motor imagery training in VR does not affect the induced event-related desynchronization: a pilot study. Front. Virtual Real. **4**, 1265010 (2024). https://doi.org/10.3389/frvir.2023.1265010. (in English)
10. Zich, C., Debener, S., Kranczioch, C., Bleichner, M.G., Gutberlet, I., De Vos, M.: Real-time EEG feedback during simultaneous EEG-fMRI identifies the cortical signature of motor imagery. Neuroimage **114**, 438–447 (2015). https://doi.org/10.1016/j.neuroimage.2015.04.020. (in English)
11. Saha, S., Ahmed, K.I.U., Mostafa, R., Hadjileontiadis, L., Khandoker, A.: Evidence of variabilities in EEG dynamics during motor imagery-based multiclass brain-computer interface. IEEE T. Neur. Syst. Rehabil. **26**(2), 371–382 (2018). https://doi.org/10.1109/Tnsre.2017.2778178. (in English)
12. Huang, G., et al.: Discrepancy between inter- and intra-subject variability in EEG-based motor imagery brain-computer interface: Evidence from multiple perspectives. Front. Neurosci. **17**, 1122661 (2023). https://doi.org/10.3389/fnins.2023.1122661. (in English)
13. Hwang, H.J., Kwon, K., Im, C.H.: Neurofeedback-based motor imagery training for brain-computer interface (BCI). J. Neurosci. Meth. **179**(1), 150–156 (2009). https://doi.org/10.1016/j.jneumeth.2009.01.015. (in English)

14. iBadia, S.B., Morgade, A.G., Samaha, H., Verschure, P.F.: Using a hybrid brain computer interface and virtual reality system to monitor and promote cortical reorganization through motor activity and motor imagery training. IEEE Trans. Neural Syst. Rehabil. **21**(2), 174–181 (2012)
15. Lotte, F., et al.: Combining BCI with virtual reality: towards new applications and improved BCI. In: Allison, B.Z., Dunne, S., Leeb, R., José, D.R., Millán, A.N. (eds.) Towards Practical Brain-Computer Interfaces, pp. 197–220. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-29746-5_10
16. Alimardani, M., Nishio, S., Ishiguro, H.: The importance of visual feedback design in BCIs; from embodiment to motor imagery learning. PloS one **11**(9), e0161945 (2016)
17. Ruffino, C., Papaxanthis, C., Lebon, F.: Neural plasticity during motor learning with motor imagery practice: review and perspectives. Neuroscience **341**, 61–78 (2017)
18. Vourvopoulos, A., Bermúdez i Badia, S.: Motor priming in virtual reality can augment motor-imagery training efficacy in restorative brain-computer interaction: a within-subject analysis. J. NeuroEng. Rehabil. **13**, 1–14 (2016). https://doi.org/10.1186/s12984-016-0173-2
19. Crasta, J.E., Thaut, M.H., Anderson, C.W., Davies, P.L., Gavin, W.J.: Auditory priming improves neural synchronization in auditory-motor entrainment. Neuropsychologia **117**, 102–112 (2018)
20. Mizuguchi, N., Sakamoto, M., Muraoka, T., Kanosue, K.: Influence of touching an object on corticospinal excitability during motor imagery. Exp. Brain Res. **196**, 529–535 (2009)
21. Wriessnegger, S.C., Brunner, C., Müller-Putz, G.R.: Frequency specific cortical dynamics during motor imagery are influenced by prior physical activity. Front. Psychol. **9**, 398909 (2018)
22. Ziadeh, H., et al.: "Mine works better": examining the influence of embodiment in virtual reality on the sense of agency during a binary motor imagery task with a brain-computer interface. Front. Psychol. **12**, 806424 (2021)
23. Doyle, L.M., Yarrow, K., Brown, P.: Lateralization of event-related beta desynchronization in the EEG during pre-cued reaction time tasks. Clin. Neurophysiol. **116**(8), 1879–1888 (2005)
24. Vagaja, K., Vourvopoulos, A.: Electrophysiological Signals of Embodiment and MI-BCI Training in VR. https://doi.org/10.5281/zenodo.8086085
25. Türk, Ö., Özerdem, M.S.: Epilepsy detection by using scalogram based convolutional neural network from EEG signals. Brain Sci. **9**(5), 115 (2019). https://doi.org/10.3390/brainsci9050115
26. Ramirez, M., McCracken, H., Grant, B., Yielder, P., Quevedo, A.J.U., Murphy, B.A.: Using machine learning to classify eeg data collected with or without haptic feedback during a simulated drilling task. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Orlando, FL, USA, pp. 769–770. IEEE (2024)
27. Stancin, I., Cifrek, M., Jovic, A.: A review of EEG signal features and their application in driver drowsiness detection systems. Sensors **21**(11), 3786 (2021)
28. Zhao, Q., et al.: Nonlinear dynamic complexity and sources of resting-state EEG in abstinent heroin addicts. IEEE Trans. Nanobiosci. **16**(5), 349–355 (2017)
29. Hossain, M.S., Alam, K., Azad, M.A., Rahman, A., Kaiser, M.S.: An Internet of Medical Things-based patient-centric healthcare system for reliable and secure COVID-19 monitoring. IEEE Internet Things J. **9**(5), 4264–4275 (2023)
30. Yin, H., Cai, W., Wang, C., Li, J.: A bayesian approach for analyzing non-stationary multivariate time series with application to brain signal decoding. Int. J. Uncert. Fuzz. Knowl.-Based Syst. **30**(2), 245–263 (2022)

Solar Irradiance Forecasting Based on Neural Networks for Sequential Data Analysis

Alvaro D. Orjuela-Cañón[1](✉), Angie L. Blanco-Cañón[2], and Mario F. Jiménez[3]

[1] School of Medicine and Health Sciences, Universidad del Rosario, Bogota D.C., Colombia
alvaro.orjuela@urosario.edu.co
[2] Engineering Faculty, Universidad Distrital Francisco Jose de Caldas, Bogota D.C., Colombia
[3] School of Engineering, Science, and Technology, Universidad del Rosario, Bogota D.C., Colombia

Abstract. Renewable energy sources for electrical power generation are increasingly in demand as the world shifts toward sustainable solutions. Among these, solar energy has become a prominent option, with solar irradiance data being widely utilized to estimate potential energy generation. However, the ongoing impacts of climate change have introduced significant challenges in accurately measuring the availability of solar resources. To address this issue, forecasting models can play a critical role in predicting the feasibility and efficiency of solar energy generation. This paper presents a comparative approach involving two distinct neural network architectures designed for forecasting solar irradiance, which serves as a key input for photovoltaic energy systems. Specifically, the study examines the performance of long short-term memory (LSTM) networks and Transformer models to determine which approach delivers superior results under the given conditions. A first scenario proposed to forecast the irradiance with information from three previous weeks, by employing windows of 24, 48, 72, 84, 96, 120, and 144 h. The second scenario used a cross-validation technique. The analysis revealed that using data from a three-day period combined with a transformer model equipped with eight attention heads produced the most accurate and reliable forecasting outcomes. This finding underscores the potential of advanced neural network architectures in optimizing solar energy forecasting and enhancing renewable energy applications.

Keywords: LSTM · Transformer · Solar Irradiance · Time Series Forecasting

1 Introduction

Nowadays, the world fights with the challenges posed by climate change, the urgency to transition towards renewable energy sources like solar and wind power has become increasingly evident. However, the inherent variability of solar and wind resources presents a significant demanding task for an effective integration into the energy grid [1]. In this context, accurate forecasting associated to solar variables holds a crucial role in optimizing energy production processes, enhancing the grid stability, and supporting the transition to a sustainable energy future [2].

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, pp. 154–164, 2025.
https://doi.org/10.1007/978-3-031-88854-0_11

Forecasting solar resources, like irradiance, involves predicting the availability and output of these renewable energy sources over various time horizons, ranging from minutes to days ahead. Such forecasts enable grid operators, energy traders, and policymakers to make informed decisions regarding energy generation, transmission, and consumption. Moreover, with the advent of climate change, understanding and predicting the behavior of solar and wind resources under changing environmental conditions have become paramount.

In recent years, advancements in meteorology, remote sensing technologies, and data analytics have revolutionized the field of renewable energy forecasting [3]. Sophisticated numerical weather prediction models, coupled with machine learning algorithms, now enable more accurate and reliable predictions of solar irradiance. Additionally, the integration of satellite imagery, ground-based sensors, and other observational data sources enhances the spatial and temporal resolution of forecasts, enabling finer-grained insights into local weather patterns and resource availability [4].

However, despite these advancements, forecasting solar resources under the current paradigm of climate change remains a complex and evolving task [5]. The increasing frequency and intensity of extreme weather events, shifting climate patterns, and other climate-related phenomena pose new challenges for renewable energy forecasting. As such, ongoing research efforts focus on developing robust forecasting methodologies that account for the impacts of climate change on solar and wind resources.

Artificial intelligence (AI) is based on bioinspired models that mimic the natural intelligence. One example of this, the neural networks are widely used for learning patterns, for example, solar irradiance. Applications of models from AI have been employed for very fast forecasting for solar irradiance [6, 7], similar approaches with wind speed have been studied, [8], and rainfall and other climatological variables, too [9, 10].

Changes in the AI area with the improvement of models that are employed with information with time dependencies, such as sequences or time series area, which was impacted in 2017 with the inclusion of the seminar paper *Attention is all you need*, allowing to analyze the sequences in a different way [11, 12]. Due to this, the use of proposed architecture called transformer is currently used in the Large Language Models, like GPT, Bard, Gemini, among others [13, 14]. However, the power of transformers can be exploited in the time series applications in different scenarios [15, 16], which depends on the specific data series, sample rate, phenomena, among other aspects. Previous to Transformers, recurrent neural networks were widely employed for forecasting tasks. One of most versatile models has been the Long Short Memory Term (LSTM) neural network, which has the property to memorize the specific moments to develop the forecasting. Applications with the use of this neural network for solar irradiance measures can be detailed in [17–19]. In a similar way, the use of other strategies based on modifications of this or other architectures have been implemented, however, the results depend on the specific problem.

Colombia is a country situated in the northwest corner of South America. This makes that climate be characterized by its diverse geography, encompassing lush rainforests, rugged mountains, expansive plains, and stunning coastline along both the Pacific Ocean and the Caribbean Sea [3, 4]. The country's varied topography gives rise to a wide range of ecosystems, including the biodiverse Amazon rainforest in the southeast, which makes that occupies a strategic location that serves as a bridge between the continents of North and South America. Additionally, phenomenon like El Niño contributes to the difficult task of obtaining information from the weather. In this way, the forecasting of renewable resources is useful to provide information oriented to establish policies and strategies for electrical energy generation.

This paper examines the proposal of comparing two different techniques based on artificial neural networks architectures employed for time series forecasting. The objective is to supply information based on predictions that support making decisions on challenges associated with the variability of solar resources in the context of climate change. For forecasting, the LSTM and Transformer architectures were analyzed, exploring similar parameters associated to the data and hyperparameters of the networks.

2 Methods

2.1 Dataset

The Colombian meteorological and climatological institute (IDEAM: *Instituto de Hidrología, Metereología y Estudios Ambientales*) holds a platform where the data from different climatological variables are available from different places of Colombia, according to acquisition its systems [20]. For this case, data from Valle del Cauca state was acquired for the study. Information with hour intervals time from January to August of 2018 was taken into account. A bigger period of time was not considered due to missing values related to acquisition issues. Figure 1 shows the first ten days of the acquisition, visualizing the solar irradiance variable.

For training of the models, the month of January was elected for the study, delimiting the complete sequence to thirty-one days, according to scenario where data acquisition is restricted as a real one. In a first scenario, the dataset was divided into a training set with the information from three weeks, and test set was adjusted to the last week of the month. This idea is supported in a possible implementation of a system that can acquire data during this period of time and a horizon of one week. The second scenario divided the information into four folds to apply a technique of cross-validation.

A normalization was implemented, bounding the values to the interval [0, 1]. The dataset division was established after the normalization, according to avoid to explode the parameters of the neural network architectures. In addition, this type of normalization is preferable in time series scenarios due to the preservation of the series behavior.

2.2 Neural Networks Architectures for Sequential Data Analysis

The LSTM neural networks are a type of recurrent neural network (RNN) architecture specifically designed to address the vanishing gradient problem and capture long-range dependencies in sequential data [21]. The traditional RNNs had problems associated to long-term in time series, LSTM networks incorporate specialized memory cells that allow them to maintain information over extended time periods. The architecture of the LSTM is based on cells, which consists of a set of three gates: input gate, forget gate, and output gate. The specific coupling of them allows to be responsible for regulating the flow of information through the cell [22]. These gates, controlled by sigmoid and tanh activation functions, determine which information to retain, discard, or update at each time step, enabling LSTMs to effectively learn and remember patterns in sequential data.

One explored hyperparameter of this architecture was the number of cells, this value was tested from one to five. The implementation was obtained by employing the Keras framework, and using a dense layer in the output, according the univariate time series. More cells were not used, due to the advantaged abilities of LSTM networks to handle sequences of varying lengths, capturing both short-term and long-term dependencies within the data. Additionally, LSTM networks have been extended and adapted in various ways to address specific challenges, leading to the development of variants such as bidirectional LSTMs, stacked LSTMs, and attention-based LSTMs, further enhancing their versatility and applicability across different domains.

Second architecture analyzed was the Transformer neural network. This kind of models comprises several key components that allow to capture and process information of sequence data in a unique way. Its architecture includes two essential parts: encoder and decoder, each one composed with multiple identical layers. Each layer holds two principal subcomponents: self-attention mechanisms and position-wise feedforward networks. Self-attention mechanisms enable the model to weigh the relevance of each word in a sequence by considering the relationships between all words simultaneously. This is achieved through queries, keys, and values, which are linear projections of the input embeddings. These embeddings are the representation of data through the vectorization of the sequence to process. This allows the parallelization of the attention, employing different number of heads. By attending to different parts of the input sequence, the self-attention mechanism allows the model to capture both local and global dependencies effectively. In a similar mode as the LSTM one, the hyperparameter of the architecture was modified, changing the value of the heads in the process following the embedding with a size of 64. The value of the heads needed to be a divisor of this number, then, values of 2, 4, 8, 16, and 32 were tested, homologous to five values in the number of cells of the LSTM.

In addition, the network has position-wise feedforward networks that consist of fully connected layers applied independently to each position in the sequence [12]. This part introduces additional flexibility and expressiveness to the model, enabling it to learn complex relationships and patterns in the data. Moreover, residual connections and layer normalization are employed to stabilize the training process and facilitate the flow of information through the network. Together, these internal components of the Transformer architecture allow for efficient processing of sequential data while capturing both short-term and long-term dependencies, leading to state-of-the-art performance across various natural language processing tasks [13].

Finally, other parameter to analyze was the period of time previous to make the forecasting. For this, it was employed a vector with a specific number of days. Due to the samples are given by hours, the vectors included 24, 48, 72, 96, 120, and 144 h. This means that use of data from one day to six days were presented to the network training, according to real applications. Two scenarios were implemented: *i)* real-application: the dataset was divided into a train set with information from first three weeks, and testing on the last week; *ii)* a cross-validation technique with four folds, where the entire dataset was divided in exact segments, different to the previous week size. According to the time dependency data, it was taken into account in the division.

2.3 Metrics

The root mean square error (RMSE) is a widely used metric in regression analysis to quantify the accuracy of a predictive model by measuring the average magnitude of the differences between predicted and actual values. This metric is calculated through the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}}.$$

where $\hat{y}$ represents the predicted values, y represents the original values, n is the number of data points, and Σ denotes summation. RMSE provides a single value that indicates the typical deviation of predicted values from actual values, with lower RMSE values indicating better model performance.

3 Results

As mentioned, two scenarios were implemented according to the validation strategy. The results for the first scenario are presented, where the training was developed with three weeks from January and validation followed a hold-out strategy with the fourth week. Then, second scenario was obtained through a cross-validation strategy where the dataset was divided into four folds. Then four trainings were implemented, employing approximately three weeks for training and the test with the fourth week. In this strategy, the test set was not of the week size exactly due to the division of the dataset. The number of hours in each dataset was used as reference for the division, where 186 h was the size for the test set compared to 168 h used in the first scenario. Figure 1 shows an example of the first ten days of the solar irradiance series.

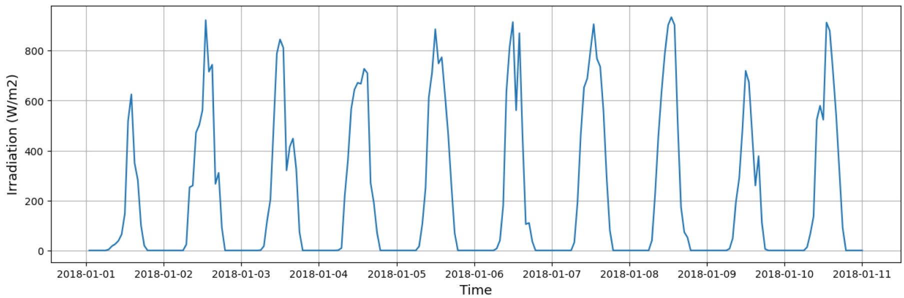


Fig. 1. Example of the solar irradiance time series values.

In the first scenario the experimentation was implemented for evaluation of lags or previous information necessary to make the forecasting. In addition, the number of units in the proposed architectures as mentioned. Tables 1 and 2 show the results for the employment of LSTM neural network and the Transformer architecture, respectively, in this scenario. The tables include the periods of time, in hours, for making possible the forecasting. Due to the number of days for the test set is seven days, the longest period could be six days, 144 h, to forecast the only missing last day.

Figure 2 and 3 show the comparison for the forecasting for both models. Figure 2 visualizes the comparison in the best scenario for the LSTM model, where the forecasting needed data from six previous days. There, the LSTM obtained the best performance with an input of 144 h and five cells in the neural network. For the Transformer architecture case, it was necessary an input of 72 h, three previous days, and three heads after the embedding. For both figures, the color blue is associated to the original test set, orange color is for the prediction of the LSTM model, and green color is for the Transformer model. In addition, for both cases, the straight line is the needed data used as input for each model and to do the forecasting.

Table 1. Results of the RMSE for the LSTM architecture

Hours	1 Unit	2 Units	3 Units	4 Units	5 Units
24	0.1762	0.1550	0.1525	0.1614	0.1587
48	0.1864	0.1760	0.1679	0.1717	0.1877
72	0.2163	0.1956	0.1791	0.2231	0.1783
96	0.2290	0.2179	0.1956	0.2207	0.2192
120	0.2646	0.2375	0.2508	0.2589	0.2454
144	0.2532	0.1816	0.1736	0.2401	**0.1514**

Table 2. Results of the RMSE for the Transformer architecture

Hours	1 Head	2 Heads	3 Heads	4 Heads	5 Heads
24	0.1183	0.1177	0.1149	0.1112	0.1128
48	0.1201	0.1112	0.1129	0.1159	0.1234
72	0.1107	0.1150	**0.1071**	0.1135	0.1086
96	0.1202	0.1220	0.1221	0.1184	0.1171
120	0.1167	0.1122	0.1172	0.1192	0.1162
144	0.1272	0.1244	0.1179	0.1301	0.1240

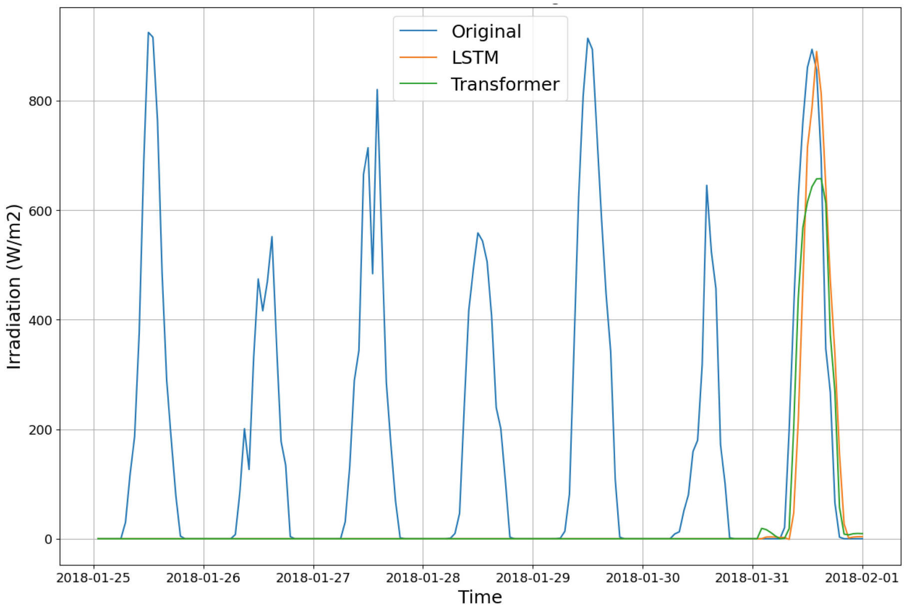


Fig. 2. Forecasting for the best performance of the networks based on LSTM architecture.

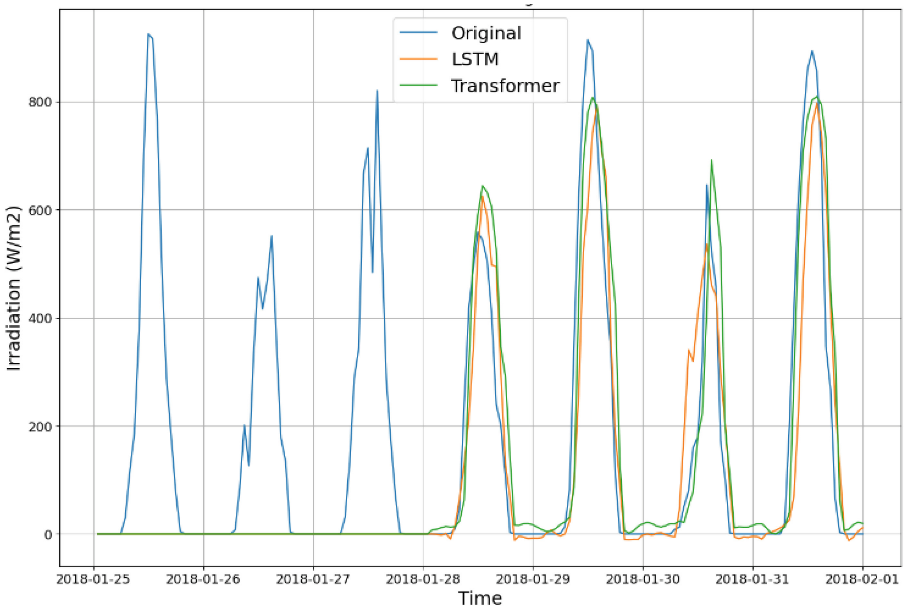


Fig. 3. Forecasting for the best performance of the networks based on Transformer architecture.

The second scenario based on cross-validation can be summed through the Tables 3 and 4. There, it is reported the best value in each fold and window input for each network. In addition, the number of units (Table 3) and heads (Table 4) where it was found the best value is exhibited. It is possible to see the differences between the values at comparing the tables. LSTM networks had higher RMSE values but with less quantity of units for the architectures compared to the Transformer networks. The minimum value for both architectures was reached for the input size of 120 h with three units for the LSTM and four heads for the Transformer.

Table 3. Results using cross-validation for the LSTM architectures

Hours	Fold 1		Fold 2		Fold 3		Fold 4	
	RMSE	Units	RMSE	Units	RMSE	Units	RMSE	Units
24	0.1400	5	0.1954	1	0.1602	2	0.1807	2
48	0.1814	2	0.1942	2	0.1517	2	0.1913	3
72	0.1699	1	0.2124	1	0.1597	1	0.2205	1
96	0.1882	1	0.2006	4	0.1520	5	0.2285	4
120	0.1979	3	0.1607	5	**0.1300**	**3**	0.2377	4
144	0.2053	3	0.1383	1	0.1504	1	0.2510	3

Table 4. Results using cross-validation for the Transformer architectures

Hours	Fold 1		Fold 2		Fold 3		Fold 4	
	RMSE	Heads	RMSE	Heads	RMSE	Heads	RMSE	Heads
24	0.1153	4	0.1297	3	0.1234	3	0.1347	5
48	0.1128	2	0.1301	1	0.1189	4	0.1363	2
72	0.1182	2	0.1291	5	0.1178	5	0.1369	3
96	0.1148	5	0.1311	4	0.1160	1	0.1384	2
120	0.1211	5	0.1206	5	**0.1065**	**4**	0.1420	1
144	0.1191	5	0.1185	4	0.1179	5	0.1319	4

4 Discussion

From the results, it is seen that the LSTM models have worse performance compared to the Transformer ones. This is appreciated in the Tables 1 and 2, where the values, in general, for the Table 1 are upper than Table 2. Other important aspect that can be observed is the quantity of looking back is necessary for each model. The LSTM proposal needs six days with data from 144 h to make the best forecasting process, as seen in

Fig. 2. Furthermore, the Transformer, with information from three days (72 h) previous to the forecasting, can develop a better forecasting process. This can be evidence of the Transformer's power and the pay attention of data [11].

In spite of the differences between the employed models, a visual inspection of the results exhibits an operation that follows the movements in the time series, associated to diary pattern that characterize the solar irradiance. In general terms, it is possible to propose any of two models for forecasting tasks.

Analyzing the second scenario, it was possible to see how the Transformer architectures had best results than LSTM ones, in a general way. From Tables 3 and 4 can be visualized the differences of this effect and the bigger number of heads contrasted to units in the RNN networks. At the same time, comparing the values for the different folds, it is possible to see how the last fold had the higher values for the RMSE. This is due to the division of the dataset, establishing the first week as the test set and three last weeks as training, which lacks of representation of a real application.

Different works based on the Colombian specific problem has been reported in [3]. There, there were used models based on fuzzy systems and artificial neural networks. However, the architecture was distant to the presented in this paper, and the results in that case were no conclusive. More details were presented in [23], where best models were found based on support vector regressor. Other Colombian studies have been reported in similar regions that the present one like Nariño, with LSTM, and other lineal models [24]. In addition, data from the Caribbean region of the country was analyzed by employing just a feedforward network [25], and more locations of Colombia were analyzed in a multivariate context with neural networks and other machine learning models [26].

However, it is important to note that the different reported works, for the Colombian case, did not employ transformers architecture within their proposals. Limitations of the present analysis is based on the variation of the hyperparameters of considered models. However, it is difficult to establish a clear comparison due to the architectures are totally different but with fundamentals in neural networks. LSTM depends on the modification of the cells in the model, meanwhile the Transformer holds many more parameters to choose. For example, number of layers in the feedforward part, size of the embedding. These parameters do not have a comparable parameter in the LSTM model. Finally, computational costs associated to the implementation for each architecture must be calculated and reported.

5 Conclusions

This present proposal visualizes the comparison between two neural networks architecture employed for time series forecasting. LSTM and Transformer models were assessed to forecast the solar irradiance in Colombia. The results showed the power of the Transformer model, which needed data from three previous days for making the best prediction.

LSTM architecture continue supporting time series forecasting, but for the analyzed case, the results were not achieved the performance of the Transformer and its structure based on attention. From different implementations modifying input sizes and units in the architectures, it was possible to establish that Transformer networks represent better alternatives to forecast the solar irradiance in the present case.

Acknowledgments. Authors acknowledge the support of the Universidad del Rosario for funding this project. In addition, the contribution of the seedhub team *Semillero en Inteligencia Artificial en Salud: Semill-IAS*. In addition, authors declare and acknowledge the use of artificial intelligence text generation system through ChatGPT in some sections of this paper.

References

1. Santhosh, M., Venkaiah, C., Vinod Kumar, D.M.: Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: a review. Eng. Rep. **2**, e12178 (2020)
2. Yang, D., et al.: A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: towards carbon neutrality. Renew. Sustain. Energy Rev. **161**, 112348 (2022)
3. Obando, E.D., Carvajal-Quintero, S.X., Agudelo, J.P.: Modeling and assessment of machine learning models for solar radiation forecast. In: 2023 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT-LA), pp. 490–494 (2023)
4. Carmona, M., Eljaik, O., Huertas, J.: The energy efficiency of autorickshaws in a coastal region of Colombia. Transp. Res. D Transp. Environ. **121**, 103802 (2023)
5. Pryor, S.C., Barthelmie, R.J., Bukovsky, M.S., Leung, L.R., Sakaguchi, K.: Climate change impacts on wind power generation. Nat. Rev. Earth Environ. **1**, 627–643 (2020)
6. Orjuela-Cañón, A.D., Hernández, J., Rivero, C.R.: Very short term forecasting in global solar irradiance using linear and nonlinear models. In: 2017 3rd IEEE Workshop on Power Electronics and Power Quality Applications, PEPQA 2017 - Proceedings (2017). https://doi.org/10.1109/PEPQA.2017.7981691
7. Tealab, A.: Time series forecasting using artificial neural networks methodologies: a systematic review. Fut. Comput. Inf. J. **3**, 334–340 (2018)
8. Men, Z., Yee, E., Lien, F.-S., Wen, D., Chen, Y.: Short-term wind speed and power forecasting using an ensemble of mixture density neural networks. Renew. Energy **87**, 203–211 (2016)
9. Rivero, C.R., et al.: Time series forecasting using recurrent neural networks modified by bayesian inference in the learning process. In: 2019 IEEE Colombian Conference on Applications in Computational Intelligence, ColCACI 2019 - Proceedings (2019). https://doi.org/10.1109/ColCACI.2019.8781984
10. Rivero, C.R., et al.: Bayesian enhanced ensemble approach (BEEA) for time series forecasting. In: 2018 IEEE Biennial Congress of Argentina (ARGENCON), pp. 1–7 (2018)
11. Vaswani, A., et al.: Attention is all you need. Adv Neural Inf Process Syst. **30** (2017)
12. Ahmed, S., Nielsen, I.E., Tripathi, A., Siddiqui, S., Ramachandran, R.P., Rasool, G.: Transformers in time-series analysis: a tutorial. Circuits Syst. Signal Process. **42**, 7433–7466 (2023)
13. Wen, Q., et al.: Transformers in time series: a survey. arXiv preprint arXiv:2202.07125 (2022)
14. Wu, N., Green, B., Ben, X., O'Banion, S.: Deep transformer models for time series forecasting: the influenza prevalence case. arXiv preprint arXiv:2001.08317. (2020)
15. Sharda, S., Singh, M., Sharma, K.: RSAM: robust self-attention based multi-horizon model for solar irradiance forecasting. IEEE Trans. Sustain. Energy. **12**, 1394–1405 (2020)
16. Wu, Y.-K., Huang, C.-L., Phan, Q.-T., Li, Y.-Y.: Completed review of various solar power forecasting techniques considering different viewpoints. Energies (Basel) **15**, 3320 (2022)
17. Heidari, A., Khovalyg, D.: Short-term energy use prediction of solar-assisted water heating system: application case of combined attention-based LSTM and time-series decomposition. Sol. Energy **207**, 626–639 (2020)

18. Sorkun, M.C., Incel, Ö.D., Paoli, C.: Time series forecasting on multivariate solar radiation data using deep learning (LSTM). Turk. J. Electr. Eng. Comput. Sci. **28**, 211–223 (2020)
19. Rajagukguk, R.A., Ramadhan, R.A.A., Lee, H.-J.: A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power. Energies (Basel) **13**, 6623 (2020)
20. Mateus, C.A.G.: Propuesta de una herramienta de análisis del clima a través de predictivas de datos. GRAINE. Boletín de Investigaciones. **2**, 13–21 (2020)
21. Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**, 2222–2232 (2016)
22. Zhao, Z., Chen, W., Wu, X., Chen, P.C.Y., Liu, J.: LSTM network: a deep learning approach for short-term traffic forecast. IET Intel. Transp. Syst. **11**, 68–75 (2017)
23. Paredes, E.D.O.: Machine learning model for primary solar resource assessment in Colombia. TecnoLógicas **26**, e2789–e2789 (2023)
24. Hoyos Gómez, L.S.: Forecasting the global solar radiation in Nariño–Colombia (2021)
25. Silvera, O.C., Chamorro, M.V., Ochoa, G.V.: Wind and solar resource assessment and prediction using Artificial Neural Network and semi-empirical model: case study of the Colombian Caribbean region. Heliyon **7**, e07959 (2021)
26. Miranda, E., Fierro, J.F.G., Narváez, G., Giraldo, L.F., Bressan, M.: Prediction of site-specific solar diffuse horizontal irradiance from two input variables in Colombia. Heliyon **7**, e08602 (2021)

Author Index

A. D. Orjuela-Cañón et al. (Eds.): ColCACI 2024, CCIS 2212, p. 165, 2025.
https://doi.org/10.1007/978-3-031-88854-0