Modelo de regresión simbólica basado en la optimización de Harris Hawks para la predicción de la temperatura en módulos fotovoltaicos bifaciales

Symbolic Regression Model Based on Harris Hawks Optimization for Temperature Prediction in Bifacial PV Modules

DOI: https://dx.doi.org/10.17981/ingecuc.22.1.2026.08
Artículo de Investigación Científica. Fecha de Recepción: 22/08/2025, Fecha de Aceptación: 17/06/2026.

Fabian Alonso Lara-Vargas https://orcid.org/0000-0001-8246-1852
Universidad Pontificia Bolivariana, Monteria, Colombia
Universitat Politècnica de València, Valencia, España
fabian.lara@upb.edu.co

Carlos Vargas-Salgado https://orcid.org/0000-0002-9259-8374
Universitat Politècnica de València, Valencia, España
carvarsa@upvnet.upv.es

Omar Pinzón-Ardila https://orcid.org/0000-0001-8765-1479
Universidad Pontificia Bolivariana, Bucaramanga, Colombia
omar.pinzon@upb.edu.co

Oscar J. Suarez-Sierra https://orcid.org/0000-0002-6754-5713
Universidad de Pamplona, Pamplona, Colombia
ejemplo@email.com

To cite this paper
F. Lara-Vargas, C. Vargas-Salgado, O. Pinzón-Ardila & O. Suarez-Sierra "Symbolic Regression Model Based on Harris Hawks Optimization for Temperature Prediction in Bifacial PV Modules," INGE CUC, vol. 22, no. 1, 2026. DOI: *https://dx.doi.org/10.17981/ingecuc.22.1.2026.08*

Resumen

Introducción: La predicción precisa de la temperatura en módulos fotovoltaicos (FV) bifaciales es esencial para optimizar la eficiencia energética y prolongar la vida útil del sistema. Este estudio introduce un modelo de regresión simbólica optimizado mediante el algoritmo Harris Hawks Optimization (HHO) y lo compara con un modelo basado en algoritmos genéticos (GA) y modelos estadísticos, utilizando datos reales de una planta FV bifacial de 26,6 MW ubicada en Colombia.
Objetivo: Desarrollar un modelo interpretable de regresión simbólica para predecir la temperatura de módulos FV bifaciales con seguidores solares, utilizando como variables de entrada la radiación solar y la hora solar.
Metodología: Se diseñaron y compararon cuatro modelos: regresión lineal múltiple (MLR), MLR con gradiente descendente (GD), regresión simbólica (SR) con GA y con HHO. Se emplearon datos medidos cada 5 minutos durante un año. Se analizaron la correlación entre variables, la normalidad de los datos y se aplicaron métricas como RMSE y R^2 para la validación.
Resultados: El modelo con MLR-GD obtuvo el mejor desempeño (RMSE: 4,92; R^2: 0,86), seguido por SR-GA (RMSE: 7,14; R^2: 0,71). El modelo SR-HHO mostró rápida convergencia con pocos datos, aunque menor precisión en grandes volúmenes (RMSE: 13,91; R^2: 0,09).
Conclusiones: Los modelos simbólicos permiten interpretar las relaciones térmicas en módulos FV bifaciales. El HHO destaca por su eficiencia con pequeños volúmenes de datos, mientras que GA ofrece mayor estabilidad. Se sugiere una estrategia híbrida que combine ambas técnicas para mejorar el rendimiento predictivo.

Palabras clave

Regresión simbólica; Optimización por halcones de Harris (HHO); Algoritmo genético (GA); Módulos fotovoltaicos bifaciales (PV); Predicción de temperatura.

Abstract

Introduction: Accurate temperature prediction in bifacial photovoltaic (PV) modules is crucial for optimizing energy efficiency and system longevity. This study presents a symbolic regression model optimized using the Harris Hawks Optimization (HHO) algorithm and compares its performance with a Genetic Algorithm (GA)-based symbolic model and statistical methods, using real-world data from a 26.6 MW bifacial PV plant in Colombia.
Objective: To develop an interpretable symbolic regression model to predict the temperature of bifacial PV modules with solar trackers, using solar radiation and solar time as input variables.
Method: Four models were designed and compared: multiple linear regression (MLR), gradient descent-enhanced MLR, symbolic regression with GA, and symbolic regression with HHO. A one-year dataset with 5-minute resolution was used. Correlation and normality analyses were conducted, and model performance was assessed using RMSE and R^2 metrics.
Results: The gradient descent-enhanced MLR model showed the best performance (RMSE: 4.92; R^2: 0.86), followed by the SR-GA model (RMSE: 7.14; R^2: 0.71). The SR-HHO model exhibited faster convergence and better performance with smaller datasets, though it showed lower accuracy with larger data volumes (RMSE: 13.91; R^2: 0.09).
Conclusions: Symbolic models are effective for interpreting thermal behavior in bifacial PV modules. HHO is computationally efficient with small datasets, while GA provides more stable performance with large datasets. A hybrid approach combining both algorithms is recommended to improve predictive performance.

Key Words

Symbolic regression; Harris Hawks Optimization (HHO); Genetic algorithm (GA); Bifacial photovoltaic (PV) modules; Temperature prediction

I. INTRODUCTIÓN

Traditional energy sources, mainly fossil fuels, account for over 75% of global greenhouse gas emissions and 90% of CO_2 emissions, worsening climate change [1]. In this context, solar energy stands out as a key renewable alternative to reduce reliance on finite resources, providing a sustainable and environmentally friendly solution [2]. This sustainable alternative offers a clean solution for meeting energy needs [3]. Notably, bifacial photovoltaic panels are gaining popularity due to their dual-surface sunlight capture, enhancing energy production [4]. Several studies have demonstrated the superior performance of these modules compared to monofacial modules [5]. Bifacial modules capture sunlight from both sides, with the front absorbing direct radiation and the back utilizing reflected light, enhancing overall electricity generation efficiency [6]. On the other hand, using solar trackers that adjust the orientation of bifacial modules to follow the sun's path successfully maximizes the capture of solar radiation, enhancing the efficiency of the PV system with energy generation [7].
Several factors influence the efficiency of bifacial photovoltaic (PV) modules, with the incident angle effect having a particularly complex impact due to its simultaneous influence on both the front and rear surfaces of the module [8]. Dust buildup on PV panels and varying environmental temperatures can negatively impact performance, leading to substantial reductions in efficiency and electricity production [9].
Determining the temperature of bifacial PV modules is critical for enhancing solar power system efficiency. Module temperature affects electrical properties and, consequently, power production [10]. Unlike conventional monofacial modules, double-sided PV panels may experience distinct heat-related characteristics due to their capacity to absorb additional reflected or dispersed light on the rear surface [6]. Several factors, including environmental conditions (ambient temperature, solar radiation, wind speed) and installation parameters, influence bifacial PV module performance by affecting their temperature and operational efficiency [8].
Accurate modeling and prediction of module temperature are crucial for enhancing the performance of bifacial PV systems. This requires comprehending the thermal dynamics of the system and integrating environmental data into predictive models [11]. Accurate module temperature modeling and forecasting are crucial for improving bifacial PV system efficiency, demanding a thorough understanding of thermal behavior and the incorporation of environmental data into predictive frameworks [12].
Previous studies have explored the power generation characteristics of bifacial solar PV systems and the impact of solar radiation on module temperature [13]. Multiple linear regression (MLR) models are implemented using large datasets incorporating diverse environmental conditions to predict the temperature of a solar module. Nevertheless, this study exclusively utilizes monofacial modules for the experiment [14]. While MLR offers simplicity and interpretability, it is often compared to more complex models like ANN, which may provide higher accuracy. Nevertheless, MLR remains valuable due to its ease of implementation [14].

Furthermore, multiple linear regression (MLR) with gradient descent is widely employed for predicting solar panel temperature due to its simplicity and efficacy in handling linear relationships among variables [15]. This approach is particularly advantageous in scenarios where the relationship between environmental factors and solar panel temperature can be approximated linearly [16]. A key limitation of MLR is its assumption of linearity between variables, which can lead to inaccuracies when relationships are non-linear, as often occurs with complex environmental interactions influencing solar panel temperature [17]. Therefore, an approach like symbolic regression is necessary to ensure model interpretability, achieve good precision, and address non-linear relationships in the data.

Symbolic regression, a genetic programming technique, has been utilized to develop accurate forecasting models for PV systems[18], which are critical due to the inherent variability in solar power generation [19]. Examining the implementation of symbolic regression in renewable energy prediction models is crucial for elucidating the complex relationships between input variables and power output [20]. The researchers introduced a hybrid model that integrates symbolic regression with genetic programming (GP) and an artificial neural network to predict the photovoltaic power output [20]. The most frequently employed algorithms in symbolic regression encompass GP [21] and neural network-based methodologies [21].

Furthermore, the Harris Hawks Optimization (HHO) algorithm is a nature-inspired metaheuristic approach that has garnered attention for its efficacy in optimizing symbolic expressions. [22]. The algorithm emulates the cooperative hunting strategy of Harris's hawks, thereby enhancing its capacity to explore and exploit search spaces effectively [23]. While the Harris Hawks optimization algorithm offers several advantages, it has limitations. The original HHO algorithm occasionally converges too rapidly and becomes trapped in local optima [24].

While GA-based symbolic regression is well-established and offers interpretability and flexibility, it can be computationally intensive and, at times, inefficient in terms of processing speed [25]. Improvements and hybrid approaches have been developed to address these limitations [26]. Conversely, HHO demonstrates the potential for efficiency and effective search capabilities [27], although its application in symbolic regression remains nascent.

Empirical studies directly comparing these methods in symbolic regression contexts would provide more definitive information regarding their relative strengths and weaknesses in real-world environments.

This study presents a systematic comparison of symbolic regression optimized using two metaheuristic algorithms, Genetic Algorithm (GA) and Harris Hawks Optimization (HHO), for temperature prediction in bifacial PV modules with solar trackers. The comparison is carried out using real data from a 26.6 MW plant in Colombia, with 5-minute resolution over a full year, using only solar radiation and solar time as inputs. Multiple linear regression (MLR) models and MLR with gradient descent serve as interpretability baselines. The central contribution is the characterization of the behavior of GA and HHO under varying data volumes and fixed temporal partition, identifying the conditions in which each algorithm achieves better performance. Performance is quantified using RMSE, MAE, and R^2 on the training sets.

The HHO-based model is benchmarked against MLR, gradient descent-enhanced MLR, and GA-based symbolic regression, evaluating performance through the root mean square error (RMSE) and the coefficient of determination (R^2). A one-year dataset with 5-minute intervals for solar irradiance, solar hours, and module temperature is utilized. Additionally, computational time, cyclical behavior, convergence rate, and solution complexity are compared between GA- and HHO-based models.

The paper is organized as follows: Section 2 examines the photovoltaic installation, data processing, design methodology of the algorithms, and the evaluation procedure. Section 3 presents the study's predictions and evaluates the results. Finally, conclusions are drawn in Section 4.

II. MATERIALS AND METHODS

This section describes the methodology used for the temperature prediction model, see Table 1. It starts by exploring the features of a bifacial solar photovoltaic facility equipped with solar trackers. The subsequent subsection describes the data processing techniques employed in this study, such as data filtering, normality, and correlation analysis. Subsequently, the symbolic regression algorithm utilizing HHO was implemented, as were the other comparison models. The following subsection delineates the evaluation methods, which quantify the model's performance through parameters such as the root mean square error (RMSE) and the coefficient of determination (R^2).

Table 1 Methodology followed in carrying out the models and the evaluations.

System features	• Location of the system. • Analyze the characteristics of the photovoltaic system.
Data processing	• Data Acquisition. • Data filtering and normality • Correlation analysis.
Algorithm design	• Development of comparative models

	• Designing the symbolic regression algorithm based on HHO • Obtaining the equation
Result	• Obtaining the data prediction
Evaluation	• Evaluation: RSME, R^2

Characteristics of the PV system

The information used for training and objective prediction was collected from measurements taken at a facility in Colombia. The site's geographical coordinates are 8° 34′ 32.52″ N and -74° 51′ 27.72″ W. The installation comprises a bifacial solar photovoltaic facility featuring trackers designed to produce electricity for Colombia's power grid. With a capacity of 26.6 MW, the plant uses bifacial panels that have the following characteristics, as can be seen in Table 2:

Table 2. Technical characteristics of bifacial panels used in photovoltaic plants

Item	Detail
Power rating	400 Wp
Module efficiency (%)	19.7%
Normal Operating Cell Temperature	25 ºC

Data processing

Data collection was conducted using various instruments: a CR 300 datalogger from Campbell Scientific, a pyranometer with 10μV/W/m² sensitivity, an EKO MS 80, and a Campbell Scientific 110PV temperature probe (accurate to ±0.2 °C from -40 to 70 °C). The data-gathering process spanned from January 1, 2023, to December 31, 2023, with measurements taken every five minutes.

The collected data encompassed solar radiation (measured in W/m²), solar time (recorded in hours and minutes), and solar module temperature (measured in degrees Celsius). The data underwent various analytical processes to generate accurate predictions, including filtering and correlation analysis. To enhance the algorithm's efficacy, several steps were taken:

- Data points with no PV power generation, scheduled maintenance periods, or missing information were eliminated.
- The analysis was limited to 6:00 and 18:00, excluding any power generation for the grid outside this interval.

Examining the relationships between solar radiation, temperature PV module, and solar time variables requires a correlation analysis. This statistical evaluation, utilizing correlation coefficients like Spearman's, sheds light on the nature and strength of the connections among these variables. Identifying significant correlations between the variables provides insights into their interactions and effects on the bifacial solar module's heating patterns. Assessing this relationship is crucial for understanding how strongly and in what direction the variables are connected. The outcome of the Anderson-Darling normality test is instrumental in selecting the appropriate correlation technique described by Equation (1).

$$AD = -N - \frac{1}{N}\sum_{k=1}^{N}(2i-1)\left[\left(lnF(Yi) + \ln\left(1 - F(Y\,N+1-j)\right)\right)\right] \qquad (1)$$

In statistical contexts, conventional notation is employed. The letter *i* indicates the *ith* observation in an ordered sample. N represents the total number of samples. Additionally, F(x) symbolizes the cumulative distribution function, which is crucial in statistical analysis. Pearson's method is appropriate when the data has a normal distribution. Spearman's method is used in cases where the data do not follow a normal distribution. The Spearman approach is utilized, with the correlation coefficient falling between -1.0 and +1.0. It's worth noting that while a correlation may exist, it might not be linear [28].

Algorithm development

In an alternative approach, MLR with gradient descent refers to applying gradient descent methods to optimize the parameters of a multiple linear regression model [29]. This approach is particularly advantageous in scenarios where the dataset is large and traditional methods, such as standard equations, are computationally expensive [30]. The gradient descent is defined by Equations (3) and (4).

$$\theta = \min\left(j(\theta)\right). \qquad (3)$$

$$\theta_i = \theta_i - \alpha * \frac{dj(\theta_0, \theta_1)}{d\theta_i} \qquad i = 0,1..D+1, \quad (4)$$

Where $j(\theta_0, \theta_1)$ is the cost function, θ_i is the set of parameters to be estimated, $\frac{dj(\theta_0,\theta_1)}{d\theta_i}$ is the partial derivative of the cost function with respect to parameter *i*, and α is the learning factor [31]. Conversely, symbolic regression based on genetic algorithms (GA) represents a computational methodology that integrates genetic algorithms with symbolic regression to identify the mathematical expressions that optimally fit a given dataset [32]. This approach leverages the evolutionary principles of genetic algorithms to explore the space of potential mathematical models to identify the most accurate and interpretable representation of underlying data relationships [21]. Figure 1 illustrates the principal aspects of this algorithm. The graph illustrates the flow of an evolutionary algorithm. It begins with an initial population of random solutions evaluated using a fitness function. The best-performing individuals are selected and combined through crossover and mutation to form a new generation that replaces the previous one. This cycle repeats until a stopping criterion is met, progressively optimizing the model's performance.

In contrast, the HHO algorithm draws inspiration from the collective hunting tactics of Harris's hawks. These raptors demonstrate remarkable group dynamics, effectively collaborating to precisely locate, envelop, isolate, and capture their quarry [27]. Depending on the prey's escaping energy, this algorithm alternates between the exploration and exploitation stages. The energy is represented as shown in Equation (5).

$$E = E_0\left(1 - \frac{t}{T}\right) \quad (5)$$

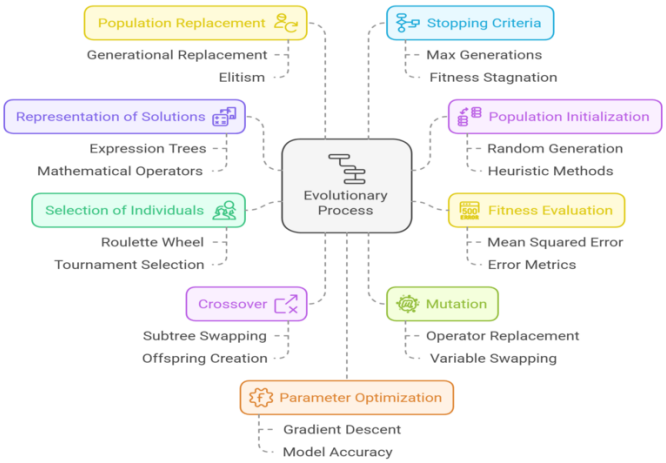

Figure 1. Diagram of the functionalities of a GA-based symbolic regression model

The variable *E* denotes the prey's escaping energy, *T* represents the maximum iteration count, and E_0 refers to the prey's energy level at the initial stage [27]. The algorithm operates in the exploration phase when *E* is greater than or equal to 1, whereas it functions in the exploitation phase when E is less than 1.

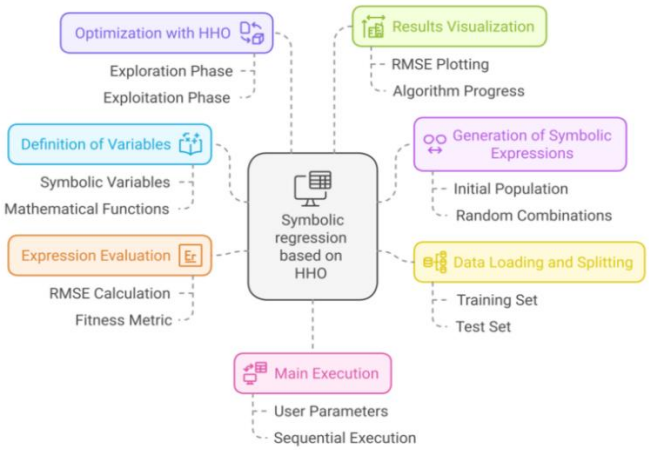

Figure 2. Diagram of HHO-based symbolic regression model functionalities

In this context, within the symbolic regression model, the primary function of the Harris Hawks Optimization (HHO) algorithm is to optimize the structure of generated symbolic expressions to identify the optimal mathematical representation that minimizes the RMSE between the model's predictions and the actual data. The above steps were undertaken to construct the model employing symbolic regression displayed in Figure 2.
Table 3 delineates the pseudocode used for constructing the symbolic regression algorithm based on HHO.

Table 3. Workflow for the construction of the symbolic regression algorithm by HHO

1. Initialization

1. Define symbolic variables (x1,x2) and available mathematical functions (e.g., \sin, \cos, \tan,\log).
2. Set constants for:
   - Population size
   - Maximum iterations
   - Number of executions
3. Load the dataset and split it into training and test sets (x1,x2,y).

2. Generate Initial Population

1. For each individual in the population:
   - Generate a symbolic expression using random combinations of:
     - Mathematical functions
     - Variables (x1,x2)
     - Random constants
   - Store the expression in the population.

3. Evaluate Population

1. For each expression in the population:
   - Evaluate its predictions for the training data.
   - Calculate its fitness using the RMSE (difference between predicted and actual values).

4. **Apply HHO Optimization**

For each iteration (t):

1. Compute energy (E=2·(1−t/max_iterations)
2. For each individual in the population:
   - If |E|≥1|E| \geq 1|E|≥1: (Exploration)
     - Replace the individual with a new random expression or use another random individual.
   - Else (|E|<1|E| < 1|E|<1): (Exploitation)
     - Perform a direct attack:
   - Replace the individual with the best solution so far.

Perform a spiral attack:
Modify the individual slightly around the best solution.
3. Re-evaluate the fitness of the population.
4. Update the best individual if a new expression achieves a lower RMSE.

5. Store Results
1. Save the best expression and its RMSE from this execution.
2. Record the time and RMSE evolution for visualization.

6. Repeat for Multiple Executions
1. For each execution:

Reset the population and repeat steps 2–5.

7. Visualization
1. Plot RMSE versus time for all executions.
2. Display the overall best symbolic expression and its RMSE.

Table 4 delineates the technical specifications of the equipment employed for developing and testing the algorithms.

Table 4. Specifications of the computing hardware employed in conducting the experiment

Item	Detail
RAM	32GB DDR4
CPU	8x1.9 GHz
RAM Speed	4267 MHz
Software	Spyder 6

In line with recommendations from prior studies, the models were trained and evaluated using a dataset split into 70% for training and 30% for testing purposes [33].

Methods of Evaluation

For the model evaluation process, the following metrics are applied:

RMSE: The process involves calculating the sum of squared differences between these values, dividing by the sample size *N*, and computing the square root of the resulting quotient. A lower RMSE value indicates higher model accuracy. Equation (6) shows the calculation method [34]:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(Vpredicted - Vtarget)^2}{N}} \quad (6)$$

R^2 (Coefficient of Determination): This measure assesses how the variance in observed values can be predicted from the model's independent variables. It is computed by subtracting from 1 the ratio of two sums: the sum of squared differences between actual and predicted values divided by the sum of squared differences between actual values and their mean. Here, *N* represents the sample size in either the calibration or validation set. Given the same concentration range, a value of *R* closer to 1 indicates superior regression or prediction performance [34]. The determination coefficient R^2 was calculated using Equation (7).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i, actual - y_i, predicted)^2}{\sum_{i=1}^{N}(y_i, actual - \bar{y}, actual)^2} \quad (7)$$

III. RESULTS AND DISCUSSIONS

This section presents the main contributions of the paper. This section presents the main contributions of the paper. The content is divided into two parts: examining the relationships between variables and evaluating model performance.

Correlation analysis

The available data for the study were adequate to develop an accurate model, as 95% of the total calculated data for a year was accessible. Before conducting the correlation analysis, a normality test was performed using the Anderson-Darling test to ensure the measured data followed a normal distribution. The calculated value exceeded the critical value

of 0.751 with a significance of 0.05, indicating that the data did not follow a normal distribution. Consequently, the Spearman correlation method was employed.

Solar radiation and the temperature of the bifacial solar module with the solar tracker have a strong correlation, as indicated by the correlation coefficient of 0.88. On the other hand, the correlation coefficient for hourly solar data was only 0.38, indicating a weaker relationship than that with solar radiation.

Under these conditions, solar time serves as a surrogate variable for the daytime environmental thermal cycle, which, at this tropical latitude, shows low interseasonal variability. This set of inputs aligns with the minimal sensor configurations assessed in recent literature on low-cost monitoring for PV systems [35].

Figure 3 illustrates the distribution of the training data. The distribution of the test data is presented in Figure 4. The relationship between the variables clearly shows non-linear patterns, especially in the daily temperature transfer.

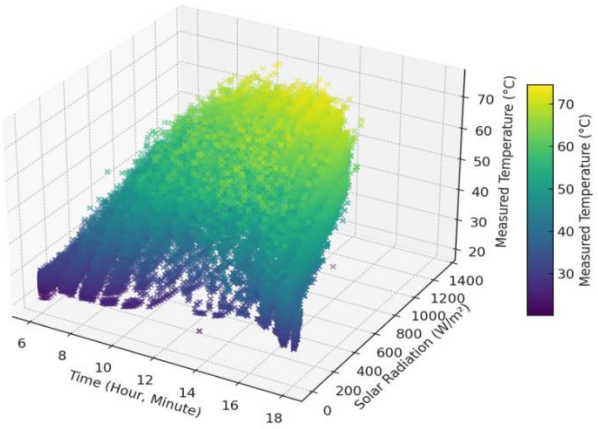

Figure 3. Relationship between time, solar radiation, and temperature measured in the training data

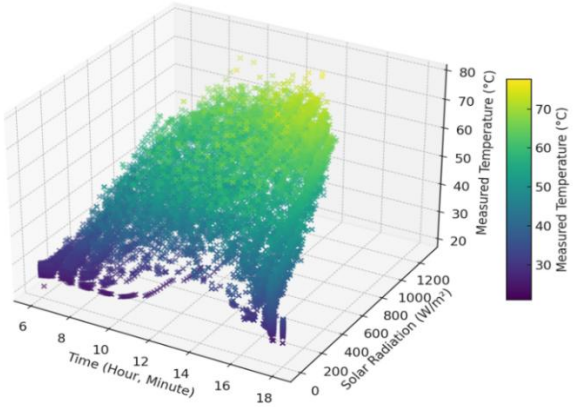

Figure 4. Relationship between time, solar radiation, and temperature measured in the test data

Evaluation of model performance

The results of the developed models are described below. First, the MLR model resulted is given by Equation (8):

$$MLR\ Model = 19.85 + 1.01x_1 + 0.03x_2 \quad (8)$$

$MLR\ GD$ Where x_1 represents the solar time expressed in hours and minutes, and x_2 represents the solar radiation in (W/m²). On the other hand, for the case of multiple linear regression with gradient descent, the result is presented in Equation (9) as follows:

$$Model = 14.5 + 2x_1 + 0.03x_2 - 0.06 * {x_1}^2 - 0.00001{x_2}^2 + 0.001x_1 * x_2 \quad (9)$$

For the development of symbolic regression algorithms based on AG and HHO, parameters are selected to enable comparison of their performances, ensuring that the population and generations/iterations are equivalent for both models. Table 5 delineates the parameters utilized in the genetic algorithm-based symbolic regression algorithm. At the same time, Figure 5 illustrates the algorithm's performance concerning the RMSE value and the execution time of each iteration.

Table 5. GA-based symbolic regression model parameters

Item	Detail
Population size	50
Number of generations	5
Number of executions	3
Mutation probability	0.5
Selection percentage	0.5
Crossover method	Subexpression Crossover

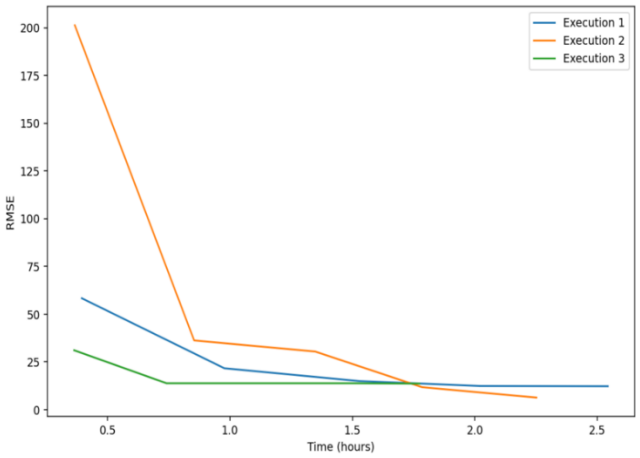

Figure 5. Evolution of RMSE behavior in different iterations SR HHO Model

The resulting equation from the execution process of the symbolic regression algorithm based on GA is described by Equation (10)

$$RS\ GA\ Model = x_1 + 1.21 * \sqrt{|x_2|} + 2atan(x_1) + 4 \qquad (10)$$

Where x_1 represents the solar time expressed in hours and minutes, and x_2 represents the solar radiation in (W/m²). The model demonstrates a non-linear relationship between solar time and solar radiation, which is consistent with the daily fluctuations of both factors and their combined influence on the temperature of bifacial photovoltaic modules. One execution of the genetic algorithm (GA) required approximately 2.5 hours, which is considered reasonable given the significant population size and data volume.

Figure 5 demonstrates an initial rapid decrease in error, particularly in Execution 2, indicating a practical initial exploration phase. However, the average initial RMSE is approximately 96. The utilization of *atan* functions may capture the daily cyclical behavior associated with solar hours. The resulting equation comprises four terms, one of which is a constant. Furthermore, for developing the symbolic regression model based on HHO, the following parameters are relevant: Table 6**Table 5** delineates the parameters utilized in the HHO-based symbolic regression algorithm, while Figure 6 illustrates the algorithm's performance concerning the RMSE value and the execution time of each iteration.

Table 6. HHO-based symbolic regression model parameters

Item	Detail
Population of Harris Hawks	50
Number of iterations	5
Number of executions	3
Decreasing energy	0-2
Direct attack	q < 0.5
Spiral attack	q >= 0.5

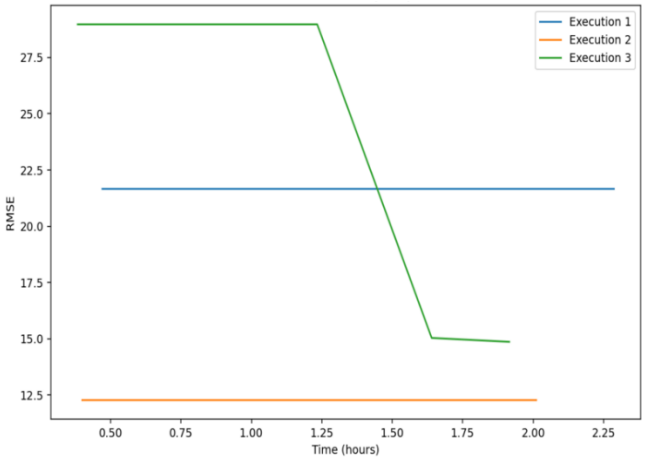

Figure 6. Evolution of RMSE behavior in different iterations SR HHO Model

The resulting equation from the execution process of the symbolic regression algorithm based on HHO is as follows, Equation (11):

$$RS\ HHO = x_1 + log(|x_1| + 1{,}05e - 5\ ) + cos(x_2) + \sqrt{|x_2|} + 2 * atan(x_1) - 0.6 \qquad (11)$$

Where x_1 represents the solar time expressed in hours and minutes, information for estimating the cosine of the angle must be provided in degrees, and x_2 represents the solar radiation in (W/m²).
This equation captures nonlinear relationships between solar time and solar radiation to model the temperature of the bifacial module. Computational time demonstrates that Execution 3 significantly reduces the RMSE from 27.5 to 13.5 after 1.75 hours. This reduction indicates an effective transition from exploration to exploitation of the search space. The HHO algorithm demonstrates an efficient approach for exploring solutions at the outset, as evidenced by the initial stability of the RMSE, with an average RMSE value below 21.
Eq. 11 contains three nonlinear terms. The term $log(|x_1| + 1{,}05e - 5\ )$ increases rapidly when $x_1 \rightarrow 0$ and levels off as the day progresses, reproducing the thermal inertia of the module in the early hours of operation; behavior consistent with the Faiman model [36], where the rate of change in temperature decreases asymptotically as accumulated irradiance increases. The term $2 * atan(x_1)$, bounded within $(-\pi, \pi)$, replicates the thermal flattening during the central hours of the day, when convective transfer offsets solar gain; its independent presence in SR-GA (Eq. 10) suggests that both algorithms identified the same functional structure to capture daytime saturation. The term $cos(x_2)$ has no direct thermophysical interpretation: within the operational range of 50 to 1000 W/m², it completes multiple cycles without associated periodicity and acts as a residual corrector bounded in [−1, 1]. Its inclusion, together with the $R^2 = -0.10$ and RMSE = 13.91 °C of SR-HHO, indicates that the algorithm fitted a local mathematical pattern at the expense of physical coherence a limitation documented in metaheuristic symbolic regression without domain constraints [21].
Table 7 presents the results of the evaluation metrics for the models using the test data. Similarly, Figure 7 Figure 8Figure 9 andFigure 10 illustrate the behavior of the models with the test data, respectively.

Table 7. Results of the evaluation metrics

Model	RMSE	R^2	MAE
MLR Model	5.29	0.83	3.90
MLR with Gradient Descent Model	4.92	0.86	3.44
SR HHO Model	13.91	0.09	12.93
SR GA Model	7.14	0.71	5.95

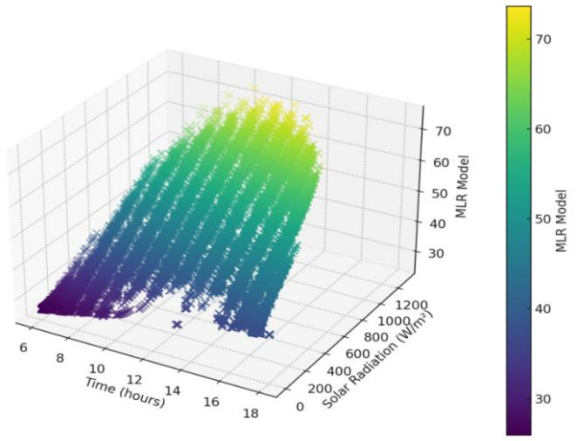

Figure 7. Evolution of RMSE behavior in different iterations (MLR Model)

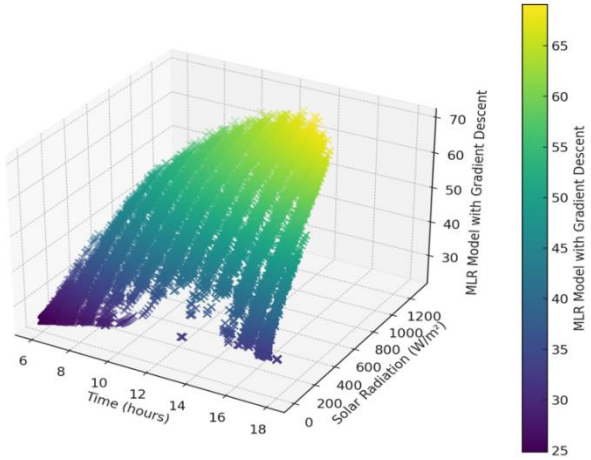

Figure 8. Evolution of RMSE behavior in different iterations (MLR GD Model)

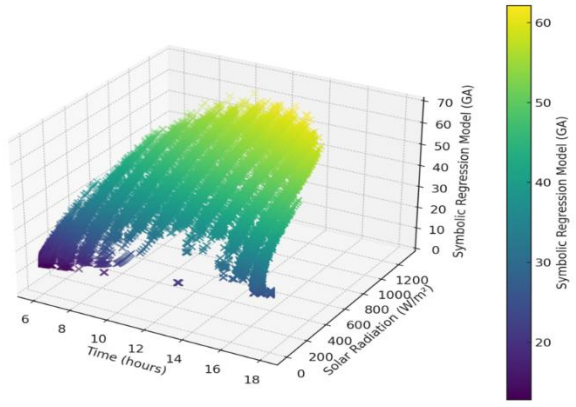

Figure 9. Evolution of RMSE behavior in different iterations (SR GA Model)

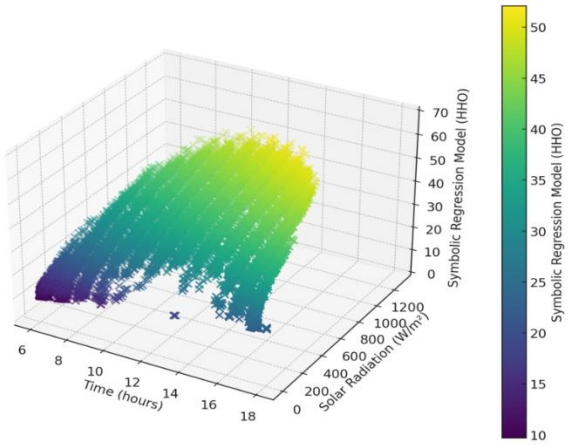

Figure 10. Evolution of RMSE behavior in different iterations (SR HHO Model)

A more rigid and linear prediction is observed in the MLR model. The MLR with Gradient Descent Model, despite having the best RMSE and R^2, exhibits voids in its surface form compared to the measured temperature in the bifacial photovoltaic solar module. The SR GA Model demonstrates superior performance under high radiation conditions. The SR HHO Model shows the lowest performance among the evaluated models.

Furthermore, to evaluate the performance of the GA and HHO-based models, a sample comprising 2.5% of the total training data was utilized to analyze the performance of these models with a reduced volume of data. To determine if there are differences in the responses of these models when dealing with varying volumes of training data. The parameters used for this test are described in Table 8 and Table 9.

Table 8. GA-based symbolic regression model parameters

Item	Detail
Population size	200
Number of generations	10
Number of executions	5
Mutation probability	0.5
Selection percentage	0.5
Crossover method	Subexpression Crossover

Table 9. HHO-based symbolic regression model parameters

Item	Detail
Population of Harris Hawks	200
Number of iterations	10
Number of executions	5
Decreasing energy	0-2
Direct attack	$q < 0.5$
Spiral attack	$q >= 0.5$

Table 10, Figure 11and Figure 12 present the evaluation results utilizing the assessment metrics and the performance of the RMSE and computational time for each algorithm, respectively.

Table 10. Results of the evaluation metrics

Model	RMSE	R^2	MAE
SR HHO Model	7.07	0.71	3.28
SR GA Model	7.63	0.69	3.41

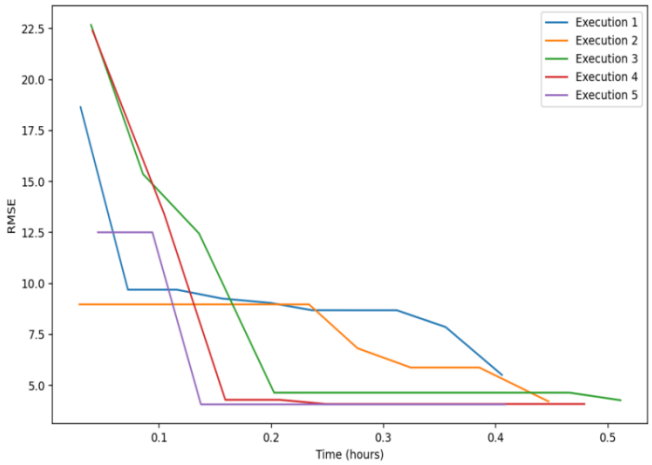

Figure 11. Evolution of RMSE behavior in different iterations RS GA Model

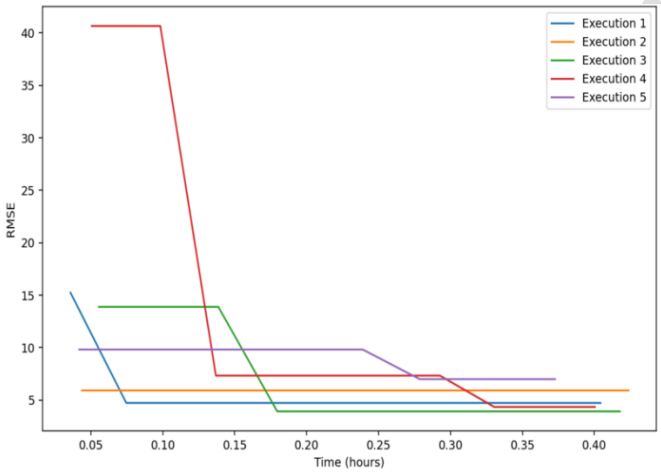

Figure 12. Evolution of RMSE behavior in different iterations SR HHO Model

As observed, the symbolic regression model based on GA requires slightly more time to complete its executions; however, the final RMSE results are comparable. Equation (12) presents the resultant equation for the model based on GA, while Equation (13) describes the model's outcome based on HHO.

$$RS\ GA\ Model = 2 * x_1 + 2 * sin(x_1) * atan\ (x_1) + \sqrt{|x_2|} \quad (12)$$

$$RS\ HHO\ Model = 2 * x_1 + log(|x_1| + 1{,}05e - 5\ ) * sin(x_2) + \sqrt{|x_2|} + 0.661 \quad (13)$$

As can be seen, the model equation based on GA contains fewer terms than the one based on HHO while simultaneously providing superior results. Table 11 compares different aspects and contexts of symbolic regression models based on GA and HHO.

Table 11. Results of the evaluation metrics

Metric	SR GA (high volume of data)	SR HHO (high volume of data)	SR GA (low volume of data)	SR HHO (low volume of data)
Convergence rate	Slow, gradual improvement	Fast, but with stagnation	Fast, but with dispersion	Very fast convergence in a few iterations
Initial RMSE	High in some runs	Lower initial variability	Lower than for large data	Lower initial variability

Final RMSE	Similar to HHO with more iterations	Similar to GA	Similar to big data	Similar to GA
Variability between executions	High in initial values	Moderate	High	Low
Computational efficiency	More computationally expensive	Better performance	Fast and efficient	Highly efficient
Complexity of the equation	Medium	High	Low	Medium

Upon analyzing the comparative performance of the GA and HHO algorithms, it is observed that HHO generally demonstrates superior overall performance, particularly in terms of convergence speed and computational efficiency. HHO is notably effective with low data volumes, achieving rapid convergence and maintaining low variability between executions. Conversely, GA exhibits a more gradual improvement and requires more significant computational resources, although it maintains consistency in the complexity of its equations.

On the other hand, it was demonstrated that GA-based methods perform effectively in equation recovery, as noted in recent literature [37].

Furthermore, it was demonstrated that the capacity of HHO to explore the search space efficiently may lead to faster convergence times compared to traditional GA methods [38]. However, this study demonstrated this capability as the core of a symbolic regression model.

Furthermore, although GA algorithms can effectively search large and complex spaces [21]. It was demonstrated that the HHO algorithm applied in symbolic regression allows for rapid search and better initial RMSE levels than GA algorithms applied in symbolic regression.

Finally, hybrid approaches could be evaluated as a complementary strategy wherein HHO is utilized initially to identify rapid solutions, and GA is subsequently employed to refine the final equation within symbolic regression.

In summary, the modeling of solar time data, solar radiation, and bifacial PV module temperature facilitated the development of a low-complexity equation with appropriate metrics for predicting the temperature of bifacial photovoltaic modules in environments with characteristics similar to those of the PV generation plant in the study.

The 70/30 random split is standard in FV temperature prediction models with independent physical inputs [39],[40]. Unlike autoregressive models, the model's inputs (solar radiation and solar time) are measured independently at each instant, so the correlation between consecutive samples does not introduce information leakage between the training and test sets. Nevertheless, an ordered temporal partition would serve as a complementary validation to assess the model's seasonal generalization.

IV. CONCLUSIONS

This study compared symbolic regression approaches optimized with Genetic Algorithms (GA) and Harris Hawks Optimization (HHO) for predicting the temperature of bifacial photovoltaic modules with solar trackers, evaluated against multiple linear regression (MLR) and MLR with gradient descent models.

The results demonstrate that, although the MLR model with gradient descent achieved the best performance in terms of RMSE (4.92) and R^2 (0.86), the GA-based model exhibited superiority under high radiation conditions due to its more stable structure and reduced susceptibility to overfitting. In contrast, the SR HHO model presented the lowest performance.

Regarding computational efficiency, HHO demonstrated rapid initial convergence, particularly with small data volumes, while GA required more computation time but achieved more consistent and accurate solutions with large data volumes. Based on these findings, it is suggested that a hybrid approach be considered, where HHO is utilized for rapid initial exploration, and GA is used for final model refinement.

Future research could further integrate additional variables, such as wind speed and humidity, to optimize temperature prediction accuracy in bifacial modules, thereby enhancing the design and efficiency of solar energy systems

V. CRediT AUTHORSHIP CONTRIBUTION STATEMENT

F. Lara-Vargas: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing-Original draft, Visualization, **C. Vargas-Salgado:** Conceptualization, Methodology, Validation, Writing-Review and editing, Supervision, Project administration, **O. Pinzón-Ardila:** Methodology, Formal analysis, Validation, Visualization, Writing-Review and editing, **O. Suarez-Sierra:** Methodology, Formal analysis, Validation, Writing-Review and editing, Supervision.

VI. FUNDING

One of the authors, F.A.L.-V., was granted a scholarship by the Universidad Pontificia Bolivariana through Act 58 of 25 October 2023 for studies at the Universitat Politécnica de Valencia.

ACKNOWLEDGMENTS

The Universidad Pontificia Bolivariana Seccional Monteria and Atlantica Colombia SA supported this work. The authors also wish to thank the Universidad Politécnica de Valencia and Universidad de Pamplona for their collaboration in developing this research.

REFERENCIAS

[1] A. I. Osman *et al.*, "Cost, environmental impact, and resilience of renewable energy under a changing climate: a review," *Environ. Chem. Lett.*, vol. 21, no. 2, 2023, doi: 10.1007/s10311-022-01532-8.

[2] A. Goswami, P. Sadhu, U. Goswami, and P. K. Sadhu, "Floating solar power plant for sustainable development: A techno-economic analysis," *Environ. Prog. Sustain. Energy*, vol. 38, no. 6, 2019, doi: 10.1002/ep.13268.

[3] M. J. H. AlTimimi, "Solar Energy," in *Quantum Dots*, J. Thirumalai, Ed., Rijeka: IntechOpen, 2022, ch. 5. doi: 10.5772/intechopen.106155.

[4] N. H. Abdul Kahar, N. H. Azhan, I. Alhamrouni, M. N. Zulkifli, T. Sutikno, and A. Jusoh, "Comparative analysis of grid-connected bifacial and standard mono-facial photovoltaic solar systems," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 4, pp. 1993–2004, Aug. 2023, doi: 10.11591/eei.v12i4.5072.

[5] A. A. B. Baloch, M. Armoush, B. Hindi, A. Bousselham, and N. Tabet, "Performance Assessment of Stand Alone Bifacial Solar Panel Under Real Time Conditions," 2018. doi: 10.1109/pvsc.2017.8366448.

[6] H. P. Yin *et al.*, "Optical enhanced effects on the electrical performance and energy yield of bifacial PV modules," *Solar Energy*, vol. 217, 2021, doi: 10.1016/j.solener.2021.02.004.

[7] L. Burnham, D. Riley, B. Walker, and J. M. Pearce, "Performance of Bifacial Photovoltaic Modules on a Dual-Axis Tracker in a High-Latitude, High-Albedo Environment," in *Conference Record of the IEEE Photovoltaic Specialists Conference*, 2019. doi: 10.1109/PVSC40753.2019.8980964.

[8] Y. Zhang, J. Q. Gao, Y. Yu, Q. Shi, and Z. Liu, "Influence of incidence angle effects on the performance of bifacial photovoltaic modules considering rear-side reflection," *Solar Energy*, vol. 245, 2022, doi: 10.1016/j.solener.2022.08.027.

[9] X. Liu, L. Cui, Q. Tao, Z. Yi, J. Li, and L. Lu, "Dust deposition mechanism and output characteristics of solar bifacial PV panels," *Environmental Science and Pollution Research*, vol. 30, no. 45, 2023, doi: 10.1007/s11356-023-29518-1.

[10] J. P. Singh, A. G. Aberle, and T. M. Walsh, "Electrical characterization method for bifacial photovoltaic modules," *Solar Energy Materials and Solar Cells*, vol. 127, 2014, doi: 10.1016/j.solmat.2014.04.017.

[11] J. Yuan, Z. Tian, J. Ma, K. L. Man, and B. Li, "A Digital Twin Approach for Modeling Electrical Characteristics of Bifacial Solar Panels," in *Proceedings - 2022 International Conference on Industrial IoT, Big Data and Supply Chain, IIoTBDSC 2022*, 2022. doi: 10.1109/IIoTBDSC57192.2022.00065.

[12] J. S. Stein, D. Riley, M. Lave, C. Hansen, C. Deline, and F. Toor, "Outdoor Field Performance from Bifacial Photovoltaic Modules and Systems," 2018. doi: 10.1109/pvsc.2017.8366042.

[13] C. Ghenai, F. F. Ahmad, O. Rejeb, and M. Bettayeb, "Artificial neural networks for power output forecasting from bifacial solar PV system with enhanced building roof surface Albedo," *Journal of Building Engineering*, vol. 56, p. 104799, Sep. 2022, doi: 10.1016/J.JOBE.2022.104799.

[14] A. Keddouda *et al.*, "Photovoltaic module temperature prediction using various machine learning algorithms: Performance evaluation," *Appl. Energy*, vol. 363, 2024, doi: 10.1016/j.apenergy.2024.123064.

[15] A. R. Kaushik, S. Padmavathi, K. S. Gurucharan, and S. C. Raja, "Performance Analysis of Regression Models in Solar PV Forecasting," in *2023 3rd International Conference on Artificial Intelligence and Signal Processing, AISP 2023*, 2023. doi: 10.1109/AISP57993.2023.10134943.

[16] P. Pourmaleki, W. Agutu, A. Rezaei, and N. Pourmaleki, "Techno-Economic Analysis of a 12-kW Photovoltaic System Using an Efficient Multiple Linear Regression Model Prediction," *International Journal of Robotics and Control Systems*, vol. 2, no. 2, 2022, doi: 10.31763/ijrcs.v2i2.702.

[17] V. Mukora, "Applying Predictive Modeling to Enhancing Solar Energy," *Virginia Journal of Business, Technology, and Science*, vol. 1, no. 2, 2022, doi: 10.51390/vajbts.v1i2.15.

[18] F. A. Lara-Vargas, J. Aguila-Leon, C. Vargas-Salgado, and O. J. Suarez, "Temperature Prediction for Photovoltaic Inverters Using Particle Swarm Optimization-Based Symbolic Regression: A Comparative Study," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 2, 2025, doi: 10.14569/IJACSA.2025.01602131.

[19] C. K. Rao, S. K. Sahoo, and F. F. Yanine, "Forecasting Electric Power Generation in a Photovoltaic Power Systems for Smart Energy Management," in *2022 International Conference on Intelligent Controller and Computing for Smart Power, ICICCSP 2022*, 2022. doi: 10.1109/ICICCSP53532.2022.9862396.

[20] M. Massaoudi, I. Chihi, L. Sidhom, M. Trabelsi, S. S. Refaat, and F. S. Oueslati, "Enhanced evolutionary symbolic regression via genetic programming for PV power forecasting," 2019.
[21] Y. A. Radwan, G. Kronberger, and S. Winkler, "A Comparison of Recent Algorithms for Symbolic Regression to Genetic Programming," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.03585
[22] Z. Liu and H. Chen, "An improved Harris Hawk optimization algorithm and its application to Extreme Learning Machine," in *2023 3rd International Conference on Consumer Electronics and Computer Engineering, ICCECE 2023*, 2023. doi: 10.1109/ICCECE58074.2023.10135354.
[23] V. Khera, "Literature Review of Harris Hawk Optimization Algorithm," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 06, no. 12, Dec. 2022, doi: 10.55041/IJSREM17248.
[24] Z. Lin, S. Jiaying, H. Chuanlu, and Z. Donglin, "Learning Harris Hawk Algorithm Based on Signal-to-Noise Ratio," *Scientific Insights and Discoveries Review*, vol. 3, pp. 236–261, Oct. 2024, doi: 10.59782/sidr.v3i1.140.
[25] G. Kronberger, F. O. de Franca, H. Desmond, D. J. Bartlett, and L. Kammerer, "The Inefficiency of Genetic Programming for Symbolic Regression -- Extended Version," *GECCO '18: Proceedings of the Genetic and Evolutionary Computation Conference*, Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.17292
[26] M. Trabelsi *et al.*, "An Effective Hybrid Symbolic Regression–Deep Multilayer Perceptron Technique for PV Power Forecasting," *Energies (Basel).*, vol. 15, no. 23, Dec. 2022, doi: 10.3390/en15239008.
[27] B. Milenković, "Implementation of Harris Hawks Optimization (HHO) algorithm to solve engineering problems," *Tehnika*, vol. 76, no. 4, 2021, doi: 10.5937/tehnika2104439m.
[28] Y. Liu, "A Short Note on Spearman Correlation: Impact of Tied Observations," *SSRN Electronic Journal*, 2017, doi: 10.2139/ssrn.2933193.
[29] R. Kalantari, K. Rahimi, and S. N. Mezajin, "Multi-Fractional Gradient Descent: A Novel Approach to Gradient Descent for Robust Linear Regression," *Engineering World*, vol. 6, pp. 118–127, Oct. 2024, doi: 10.37394/232025.2024.6.12.
[30] N. I. Norddin, M. R. Mohd Ali, N. H. Fadhilah, N. Atikah, A. Shahida, and N. H. Nohd Noh, "Multiple Linear Regression Model of Rice Production using Conjugate Gradient Methods," *MATEMATIKA*, 2019, doi: 10.11113/matematika.v35.n2.1180.
[31] E. Cuevas, O. Avalos, P. Diaz, A. Valdivia, and M. Perez, *Introducción al machine learning con Matlab*. Bogota: Marcombo, 2021.
[32] N. Jiang and Y. Xue, "Racing Control Variable Genetic Programming for Symbolic Regression," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.07934
[33] D. Kinaneva, G. Hristov, P. Kyuchukov, G. Georgiev, P. Zahariev, and R. Daskalov, "Machine Learning Algorithms for Regression Analysis and Predictions of Numerical Data," in *HORA 2021 - 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, 2021. doi: 10.1109/HORA52670.2021.9461298.
[34] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," 2022. doi: 10.5194/gmd-15-5481-2022.
[35] F. A. Lara Vargas, J. de la Ossa Rivera, S. J. Mira, C. Vargas Salgado, J. A. Leon, and E. Villabon Lopez, "Real-Time Monitoring of Solar Photovoltaic Power Plants: A Concentrated Validation Study," in *2024 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI)*, IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/ColCACI63187.2024.10666546.
[36] E. Barykina and A. Hammer, "Modeling of photovoltaic module temperature using Faiman model: Sensitivity analysis for different climates," *Solar Energy*, vol. 146, pp. 401–416, Apr. 2017, doi: 10.1016/j.solener.2017.03.002.
[37] W. La Cava *et al.*, "Contemporary Symbolic Regression Methods and their Relative Performance," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.14351
[38] V. Vardumyan Arman and K. Hakob T, "A Harris Hawks Optimizer Based Hybrid Algorithm for Optimal Design of Analog Integrated Circuits," in *Proceedings - 2022 International Conference on Frontiers of Communications, Information System and Data Science, CISDS 2022*, 2022. doi: 10.1109/CISDS57597.2022.00031.
[39] F. A. Lara-Vargas, C. Vargas-Salgado, J. Águila-León, and D. Díaz-Bello, "Optimizing Bifacial Solar Modules with Trackers: Advanced Temperature Prediction Through Symbolic Regression," *Energies (Basel).*, vol. 18, no. 8, p. 2019, Apr. 2025, doi: 10.3390/en18082019.
[40] M. Aliwi, S. Aslan, and S. Demirci, "Firefly Programming for Symbolic Regression Problems," in *2020 28th Signal Processing and Communications Applications Conference, SIU 2020 - Proceedings*, 2020. doi: 10.1109/SIU49456.2020.9302201.

Fabian Alonso Lara Vargas received Full professor of the School of Electronic Engineering at the Universidad Pontificia Bolivariana of Colombia. He obtained a degree in electronic engineering from the Universidad Pontificia Bolivariana of Colombia in 2004, a specialization in industrial control and instrumentation from the same University in 2009, a master's degree in computer project management from the University of Pamplona of Colombia in 2016 and is a PhD student in design, manufacturing and industrial project management at the Universitat Politécnica de Valencia in Spain. He is currently a professor at the Universidad Pontificia Bolivariana, Campus Monteria, Colombia. His research interests include genetic algorithms and symbolic regression, floating solar PV and energy storage. https://orcid.org/0000-0001-8246-1852

Carlos Vargas Salgado, PhD in Industrial and Production Engineering from the Polytechnic University of Valencia. Professor of the Department of Electrical Engineering of the UPV. Since 2005, he has been linked to the Institute of Energy Engineering of the UPV, where projects have been carried out in the field of renewable energies (isolated and grid-connected microgrids based on photovoltaic, wind, and biomass gasification with battery storage). https://orcid.org/0000-0002-9259-8374

Omar Pinzon Ardila, PhD in Industrial Automation & Informatics Engineering Electrical Engineer & Specialist in Advanced Engineering Studies Associate Professor, Department of Electrical Engineering, Pontificia Universidad Bolivariana (Sectional Bucaramanga) and Junior Researcher of Minciencias. Research focuses on robust/control-predictive strategies, power quality/compensation technologies (STATCOM, HVDC), and optimization of meshed hybrid AC/DC microgrids with smart transformers. https://orcid.org/0000-0001-8765-1479

Oscar J Suarez, PhD in Electrical Engineering (Cinvestav, Mexico); Electronic Engineer with specialization in Learning Strategies; Associate Professor of Mechatronics-Electrical Engineering at Universidad de Pamplona, Colombia; tenured faculty member since 2020. President of the IEEE Computational Intelligence Society Chapter Colombia (2024–25) and an ICACIT international accreditation evaluator. Research focuses on automatic control, neural-network modelling, complex-network dynamics, and applied computational intelligence. https://orcid.org/0000-0002-6754-5713