XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

An AutoML framework for Transfer Learning applied to solar radiation prediction

Samuel da Costa Alves Basílio[1,2],Leonardo Goliatt[2]

[1]*Department of Computing, Federal Center for Technological Education of Minas Gerais*
Rua José Peres, 558, Leopoldina, 36700-000, MG, Brazil
[2]*Computational Modeling Program, Federal University of Juiz de Fora*
Rua José Lourenço Kelmer, Juiz de Fora, 36036-900, MG, Brazil
samuel@cefetmg.br, leonardo.goliatt@ufjf.br

Abstract. This paper proposes a framework based on Automated Machine Learning (AutoML) combined with Transfer Learning techniques to forecast solar radiation, particularly for locations without extensive meteorological data. The objective is to enable the prediction of solar radiation values in an optimized way in cities where there is no meteorological history. The framework enables knowledge transfer from regions with abundant data to those lacking measurement histories by leveraging historical data on solar radiation, meteorological conditions, and geographic features. This approach enables the generation of reliable forecasts even in data-scarce cities, without requiring extensive historical records. A case study was conducted in Juiz de Fora, Minas Gerais, Brazil. The results show that the use of Transfer Learning alongside AutoML offers a promising strategy for enabling cities with limited measurement history to achieve accurate predictions.

Keywords: autoML, Machine Learning, Solar Radiation, Transfer Learning

1 Introduction

Distributed generation (DG) of electricity from photovoltaic solar sources has stood out in Brazil as a sustainable and decentralized alternative to energy supply. According to data from ANEEL and ABSOLAR [1, 2], 99.26% of the installed power in microgeneration (up to 75 kW) and mini-generation (above 75 kW up to 5 MW) distributed in the country comes from photovoltaic solar sources, demonstrating the broad predominance of this technology in the sector. Currently, Brazil has more than 3.3 million solar systems connected to the grid, benefiting approximately 4.96 million consumer units through the Electric Energy Compensation System.

Minas Gerais is the second Brazilian state with the largest installed capacity in distributed solar generation, totaling 4,609.5 MW, representing 12.4% of the national capacity. These demonstrate Minas Gerais' leadership in the distributed generation scenario, resulting from state public policies, investment attractiveness, and broad public engagement in the energy transition. Juiz de Fora, located in the Zona da Mata region, is Minas Gerais's tenth largest producing city, with 39.3 MW installed. It is one of the largest cities in the region economically and in terms of population size. It has an INMET[3] meteorological facility in operation since 2007, capturing different meteorological information, which will be detailed later.

Furthermore, Brazil's photovoltaic solar energy production sector has 153 manufacturers of photovoltaic kits registered with BNDES' FINAME and other manufacturers of inverters, solar trackers, modules, batteries, and string boxes. In terms of consumption profile, the residential segment accounts for most of the installed power (49.45%), followed by the commercial and services (28.45%) and rural (13.63%) sectors, reflecting the leading role of individual consumers and small businesses in adopting the technology.

As such, Minas Gerais is one of the leading national hubs for distributed solar generation, contributing significantly to the expansion of renewable energy in Brazil.

Automated Machine Learning (AutoML) emerged as a response to the growing demand for machine learning solutions that can automate critical steps in the machine learning process, such as model selection, feature engineering, hyperparameter optimization, and cross-validation. AutoML has many applications, including time series forecasting in finance, AI-assisted medical diagnostics, fraud detection, customer segmentation, and industrial process optimization[4].

Transfer Learning is a machine learning technique that leverages the knowledge acquired by a previously

trained model on one task to improve its performance on another, often related task, but with different or scarce data[5]. This approach is especially useful when the data for the specific problem is limited, as it allows reusing generalized representations that already capture relevant features of the original domain. Transfer Learning applications are vast, ranging from natural language processing to computer vision. Among the main benefits are reduced training time, lower computational costs, and the possibility of achieving better results even with smaller datasets[6].

This work presents a new autoML framework that, in addition to traditional hyperparameter optimization of machine learning models, selects the best dataset for applying Transfer Learning to solar radiation prediction. The framework selects the best machine learning model among a predefined set, optimizes its hyperparameters, performs feature selection, and chooses the best dataset to serve as a training base in the Transfer Learning process.

These results are primarily used in the implementation of new photovoltaic plants.This type of plant requires efficient prediction of solar radiation data to control the load generated on the grid. When a new plant is installed, depending on its location, there is no historical record of meteorological data for the region. Therefore, using Transfer Learning, the generated models can predict solar radiation data sufficiently well until a database is effectively stored.

This paper is organized as follows: In Section 2, we describe the materials and methods used, presenting the database and variables used. We also present the performance metrics and briefly describe the machine learning models used. This section also details the feature selection processes and the optimization approach. In Section 3, we describe the testing methodology and present the results. Finally, in Section 4, we present our conclusions, along with the possible implications of the work presented.

2 Materials and Methods

2.1 Data

The National Institute of Meteorology (*Instituto Nacional de Meteorologia* - INMET) [3] maintains and makes available data from more than 500 meteorological stations throughout Brazil, 68 of which are automatic meteorological stations in **M**inas **G**erais - MG. This work utilized data from various periods, spanning from December 2002 to December 2021, collected from 67 different stations in 66 cities across all regions of Minas Gerais. We also used data from stations in other states besides Minas Gerais to develop the methodology proposed in this work. To select these stations, we calculated the 10 closest stations for each of the 66 cities in Minas Gerais. All cities on this list were included in the database for this work, regardless of whether they are located in Minas Gerais or not. Data from 67 stations outside Minas Gerais were used: 17 from the state of **S**ão **P**aulo (SP), 13 from **R**io de **J**aneiro (RJ), 12 from **E**spírito Santo (ES), 11 from **Ba**hia (BA), 9 from **Go**iás (GO), 4 from the **D**istrito **F**ederal (DF), and 1 from **M**ato Grosso do **S**ul (MS). The availability of data at each station depends on several external factors, including the installation date, potential failures, and scheduled maintenance periods. Specifically for the Juiz de Fora station, measurements were taken between May 26, 2007, and December 31, 2021, totaling 5,333 calendar days. After accounting for outliers and missing data, the total number of entries for the Juiz de Fora station was 5,055 days. Figure 1 highlights all cities on the map where there are analyzed weather stations, with colors differing between states.

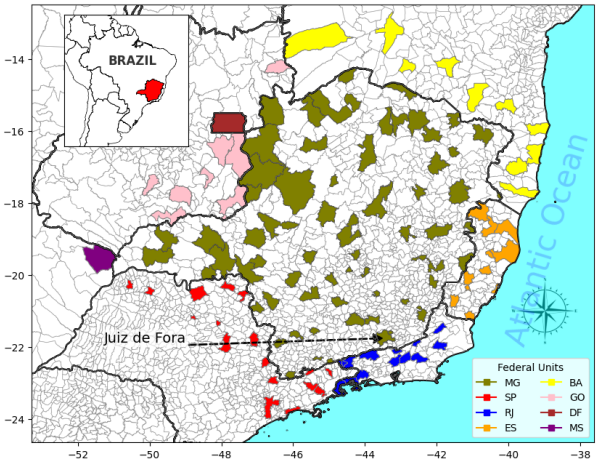


Figure 1. Map from studied municipalities.

In this work, 20 variables were used. The maximum, minimum, and average temperature for a day; The maximum, minimum, and average relative humidity of the air for a day; The maximum, minimum, and average dew point; The maximum, minimum, and average atmospheric pressure in a day; The measure of the average wind speed in a day; The average maximum wind gust speed in a day; The daily extraterrestrial solar irradiance. It is the energy available in a day at a given latitude without considering the dampening effects of the atmosphere; Latitude, Longitude, Altitude, and Biome; The measure of all solar irradiation that reaches the Earth's surface. This is the target variable of this work. Here the unit used is $MJm^{-2}d^{-1}$.

It is worth noting that the INMET data have an hourly measurement granularity. In this work, we used the daily values for each variable, considering that for the variables $Global$ and Qo, the sum of each day was used, and for the others, the average of the daily readings. The Qo variable, unlike the others, is not measured by devices and is not available through INMET. This variable was synthetically generated by the PySolar [7] library. The source of the values used for the Biome variable was provided by the Brazilian Institute of Geography and Statistics [8].

2.2 Methodology

Performance Metrics

The metrics described in Table 1, as outlined in the scikit-learn library documentation [9], were used to evaluate the performance of the methods. This Table also describes the purpose of using each metric, the best and worst results that can be found, and what they mean.

Table 1. Performance Metrics and Indicators for Model Evaluation. For each performance metric, $\hat{y}_i$ represents the predicted output, y_i is the measured output, $\bar{y}$ denotes the mean of the measured values, and N is the total number of samples in the dataset.

Performance Metric	Expression	Description	Best Result	Worst Result
R^2	$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$	Measures the proportion of variability in the dependent variable explained by the independent variables	1 (perfect fit)	0 (no fit)
RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	Measures the average deviation between observed and predicted values	0 (no error)	Higher values possible
MAE	$\frac{1}{N}\sum_{i=1}^{N}\|y_i - \hat{y}_i\|$	Measures the average deviation between observed and predicted values	0 (no error)	Higher values possible

Machine Learning Models

Below we briefly describe the machine learning models used in this work. Other models can be used by the methodology described here. The choice of these specific models was made because previous works had demonstrated their efficiency in this type of implementation.

- **Extrme Learning Machine** (ELM) is a training strategy for single-hidden layer feedforward neural networks[10]. This algorithm assigns random values to the weights of the input layer, the hidden layer, and the biases in the hidden layer. These parameters are frozen during the training process. The non-linear activation functions in the hidden layer provide non-linearity to the system. With this, the process can be considered as a linear system. The only parameter that needs to be discovered is the weight between the hidden layer and the output layer. Thus, ELM converges much faster than traditional algorithms because it learns without iteration while random hidden nodes provide a generalized approximation capability.
- **Multivariate Adaptive Regression Splines** (MARS) is a flexible regression method that automatically searches for interactions and nonlinear relationships[11, 12]. The model builds this relationship from a set of coefficients and basis functions derived from the regression data. In a way, the method is based on the 'divide and conquer' strategy, which divides the input space into regions, each with its regression equation.
- **eXtreme Gradient Boosting** (XGB) implements a boosting technique where new trees are added sequentially to correct the errors of previous trees, resulting in an increasingly robust and accurate predictive

model[13, 14]. Its ability to handle different types of data, process parallelization, and internal optimizations to reduce overfitting make it a powerful tool for classification, regression, and ranking problems.

Transfer Learning

In machine learning, Transfer Learning is a technique that leverages knowledge from a model trained on one task to enhance performance on a related, yet different, task.

Typically, the Transfer Learning process uses four steps: Pre-training, where an existing model is first trained on a large dataset for a ”source” task. This initial training phase is usually resource-intensive and takes a considerable amount of time. Feature Extraction, where the pre-trained model is then used to extract useful features or representations from the source data. These features encapsulate important patterns. Fine-tuning, where the extracted features provide a starting point for a new model designed for the target task. This new model is trained on a smaller, specific dataset for the target task. Finally, adaptation involves further training the new model on the target dataset, often with a reduced learning rate. This step enables the model to precisely adjust its learned features and optimize its performance for the unique requirements of the new task[15].

Specifically, in this work, the Transfer Learning process was employed to generate a machine learning model trained on data from weather stations in cities other than the one for which we want to make predictions. Thus, no data from the Juiz de Fora database was used in the training step, while only data from the Juiz de Fora database was used in the testing step. Figure 2 illustrates this process.

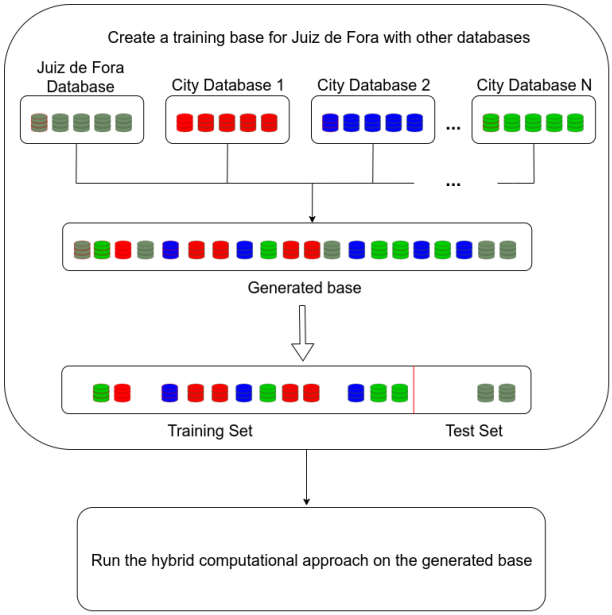


Figure 2. Transfer Learning geolocated database creation.

Automatic Feature Selection

The feature selection process is crucial in building machine learning models. This process implies selecting the most relevant variables or features from a dataset. It is essential for reducing dimensionality, enhancing model interpretability, and improving overall performance.

The developer can perform the variable selection process arbitrarily, where the variables judged attractive are used and others discarded[16]. In this process, the correlation matrix of the variables is usually observed. Considering the numerous databases and their combinations, this approach would be highly time-consuming and inefficient for the present work[17]. In this way, the feature selection process was performed automatically by an optimization algorithm, which also optimized the hyperparameters of the machine learning model. Therefore, all available variables were initially considered usable in training the final model. However, the optimization process selected only those that performed best.

A unique characteristic in this work is that the feature selection technique was also employed to select the cities that should comprise the training base in the Transfer Learning process. Thus, the optimizer, in addition to selecting the most suitable variables and the best hyperparameter values for the machine learning models, also selects which cities will generate the best results when included in the training dataset.

Hybrid Computational Approach

The methodology for developing the individual model of the city of Juiz de Fora has already been used in other works[18–20] and is described below: Initially, each machine learning model is executed individually to optimize its hyperparameters. Each model has defined upper and lower bounds for these hyperparameters. At this initial stage, we can choose to use a specific subset of variables, all available variables, or apply a feature selection technique. In this work, we employed an automatic feature selection process. After the initial settings, the optimization algorithm randomly generates a population of candidate solutions. Each solution represents a set of hyperparameters for a model, and if feature selection is automatic, each solution will have its own set of variables. Each solution is evaluated using k-fold cross-validation, to minimize the root mean square error (RMSE) between observed and predicted values. Since we are dealing with time series, we use Time Series Split Cross-Validation, ensuring future values are not used for training during the folds. The evolutionary cycle ends when the stopping criterion is met, and the solution with the lowest RMSE is stored. Finally, we analyze which model performed best and identify its optimal hyperparameter configuration. Figure 3 illustrates this process.

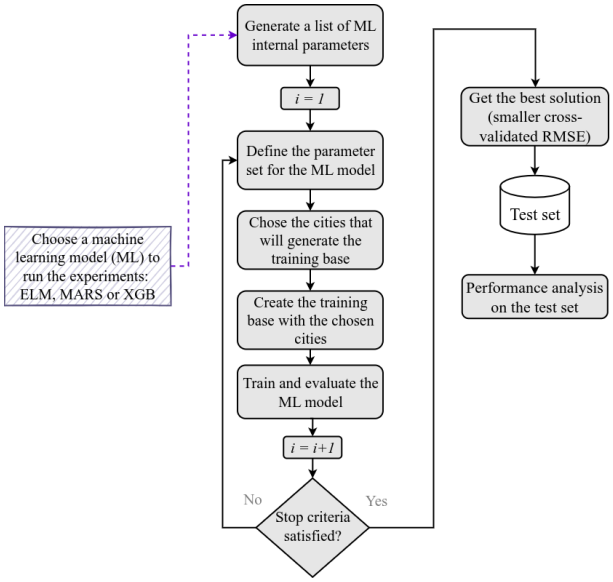


Figure 3. Optimization process.

3 Results

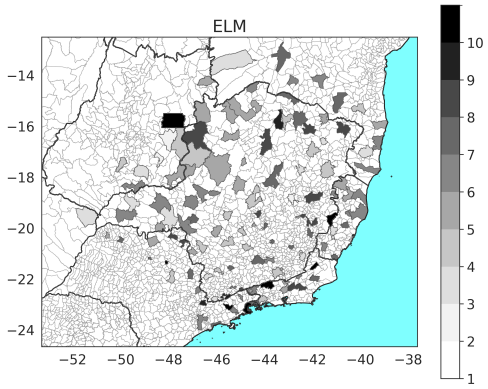
To validate the results of the Transfer Learning process, two scenarios were evaluated for each machine learning model. In both scenarios and for each machine learning model, 10 independent runs were performed. The optimizer used was a simple genetic algorithm[21] configured with 30 generations, 95% crossover factor with single-step strategy, population size of 20, and 20% mutation chance with uniform strategy.

In the first scenario, the learning model's hyperparameters were optimized without any changes to the database. The goal of this scenario is to have a benchmark for evaluating the quality of the Transfer Learning process. Only data from the Juiz de Fora station were used in both the training and testing periods. The training/testing split was 70-30. The second scenario used the previously described Transfer Learning methodology. This scenario simulates the real-world challenge of predicting solar radiation in locations with limited or no historical measurement data. The results in Table 2 show the average of 10 independent runs for each scenario and learning model. We can see that the results of the machine learning process, as expected, are inferior to those of the traditional scenario when we use the same training and testing datasets. However, the results of the Transfer Learning models are still sufficiently good when compared to those found in the literature[20].

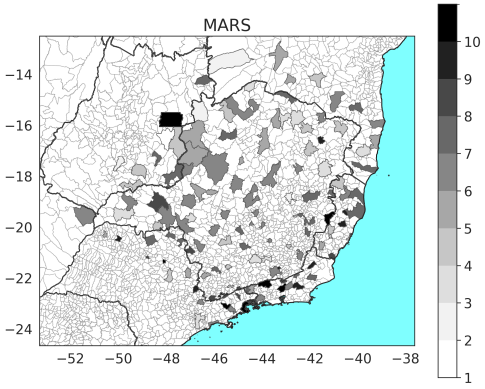
An important fact for analyzing the results is the pattern of the cities selected by the optimizer. Considering the ten independent runs, all 67 cities were selected at least once. On average, 15.7985 cities were selected in each run, with a standard deviation of 4.7673. This automatic, data-driven selection process for optimal source datasets is a key differentiator of our AutoML framework, underscoring its ability to prepare the most relevant training environment for Transfer Learning applications. Figure 4 shows the cities map, coloring the most used cities darker.

Table 2. Average Performance Metrics for Machine Learning Models in Single Base and Transfer Learning Scenarios.

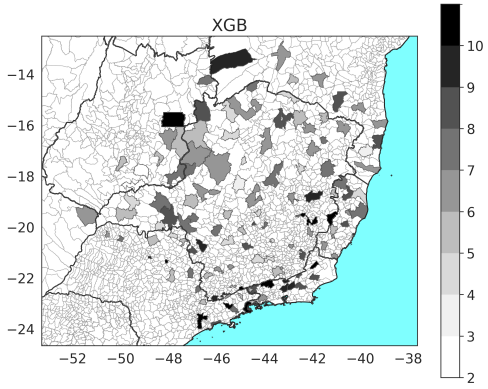
Machine learning model	Scenario	R^2	RMSE	MAE
MARS	Single base	0.867	2.22	1.73
MARS	Learning transfer strategy	0.749	3.10	2.49
XGB	Single base	0.801	2.71	2.11
XGB	Learning transfer strategy	0.739	3.15	2.53
ELM	Single base	0.826	2.54	2.00
ELM	Learning transfer strategy	0.715	3.29	2.69



(a) ELM most used cities.



(b) MARS most used cities



(c) XGB most used cities

Figure 4. Selected cities count.

4 Conclusion

This paper presents and tests a Transfer Learning methodology for predicting solar radiation. A case study is conducted using the city of Juiz de Fora as a target, along with a database containing data from 134 cities. The primary question addressed in this work is whether the Transfer Learning process can be effective for solar radiation prediction. The results show that, when compared to models using data from the target city, the Transfer

Learning process performs worse, as expected; however, it still achieves satisfactory results, consistent with the existing literature.

As future work, we suggest applying this methodology to other cities to replicate the results found in Juiz de Fora in other locations. Other machine learning models may also produce different results.

References

[1] Agência Nacional de Energia Elétrica. Sistema de Informações de Geração da ANEEL, 2025.

[2] Associação Brasileira de Energia Solar Fotovoltaica. Energia Solar Fotovoltaica no Brasil, 2025.

[3] Instituto Nacional de Meteorologia, 2022.

[4] L. Goliatt, H. A. Marhoon, Z. M. Yaseen, S. Heddam, A. W. A. Zand, B. Halder, M. L. Tan, Z. Sa'adi, I. Ahmadianfar, and S. Elsayed. An evolutionary optimized automated machine learning approach to soil unconfined compressive strength prediction for sustainable transportation infrastructure. *Transportation Geotechnics*, vol. 52, pp. 101550, 2025.

[5] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, vol. 109, n. 1, pp. 43–76, 2021.

[6] H. Echternacht, L. Campos, de A. D. Martinho, D. P. M. d. Souza, de R. B. Santis, T. S. Gontijo, M. Bodini, A. Gorgoglione, C. M. Saporetti, and L. Goliatt. Convolutional neural networks with transfer learning for natural river flow prediction in ungauged basins. *Scientific Reports*, vol. 15, n. 1, pp. 23873, 2025.

[7] Pysolar: staring directly at the sun since 2007, 2008.

[8] Instituto Brasileiro de Geografia e Estatística. Biomas do Brasil, 2019.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] J. Wang, S. Lu, S.-H. Wang, and Y.-D. Zhang. A review on extreme learning machine. *Multimedia Tools and Applications*, vol. , 2021.

[11] J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, vol. , pp. 1–67, 1991.

[12] M.-Y. Cheng and M.-T. Cao. Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. *Applied Soft Computing*, vol. 22, pp. 178 – 188, 2014.

[13] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie. Extreme gradient boosting (xgboost) model to predict the groundwater levels in selangor malaysia. *Ain Shams Engineering Journal*, vol. 12, n. 2, pp. 1545–1556, 2021.

[14] T. Chen and C. Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[15] H. H. Migaybil and B. Gopaluni. A performance neural network model for conventional solar stills via transfer learning. *Applied Energy*, vol. 375, pp. 124118, 2024.

[16] H. Hissou, S. Benkirane, A. Guezzaz, A. Beni-Hssane, and M. Azrour. *Comparative Analysis of Solar Radiation Forecasting Models: Unveiling the Potential of Feature Selection Techniques*, pp. 97–105. Springer Nature Switzerland, Cham, 2024.

[17] M. K. Boutahir, A. O. Alaoui, Y. Farhaoui, M. Azrour, and A. El Allaoui. Efficient solar radiation prediction through adaptive neighborhood rough set-based feature selection in meteorological streaming data in errachidia, morocco. In *Internet of Things and Big Data Analytics for a Green Environment*, pp. 253–272. Chapman and Hall/CRC, 2024.

[18] S. d. C. A. Basílio, C. M. Saporetti, Z. M. Yaseen, and L. Goliatt. Global horizontal irradiance modeling from environmental inputs using machine learning with automatic model selection: A case study in tanzania. *Environmental Development*, vol. , pp. 100766, 2022a.

[19] S. d. C. A. Basílio, R. O. Silva, C. M. Saporetti, and L. Goliatt. Modeling global solar radiation using machine learning with model selection approach: A case study in tanzania. In S. Shakya, K. Ntalianis, and K. A. Kamel, eds, *Mobile Computing and Sustainable Informatics*, pp. 155–168, Singapore. Springer Nature Singapore, 2022b.

[20] S. d. C. A. Basílio, F. F. Putti, A. C. Cunha, and L. Goliatt. An evolutionary-assisted machine learning model for global solar radiation prediction in Minas Gerais region, southeastern Brazil. *Earth Science Informatics*, vol. , 2023.

[21] F. Biscani and D. Izzo. A parallel global multiobjective framework for optimization: pagmo. *Journal of Open Source Software*, vol. 5, n. 53, pp. 2338, 2020.