IEEE Open Access Journal of
Power and Energy

Received 29 September 2025; revised 27 April 2026; accepted 10 May 2026. Date of publication 19 May 2026; date of current version 28 May 2026.

Digital Object Identifier 10.1109/OAJPE.2026.3694970

A Decision-Analytics Framework Combining DEA, SFA, and ML With Mixture Optimization for Industrial Photovoltaics

GUSTAVO S. LEAL[1], VICTOR E. M. VALÉRIO[1], PEDRO P. BALESTRASSI[1], LUPÉRCIO F. BESSEGATO[2], AND FARID MELGANI[3] (Fellow, IEEE)

[1]Institute of Industrial Engineering, Federal University of Itajubá, Itajubá 37500-903, Brazil
[2]Institute of Exact Science, Federal University of Juiz de Fora, Juiz de Fora 36036-900, Brazil
[3]Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy

CORRESPONDING AUTHOR: P. P. BALESTRASSI (pedro@unifei.edu.br)

This work was supported in part by Minas Gerais State Research Support Fund (FAPEMIG), in part by the National Council for Scientific and Technological Development (CNPq), in part by the Coordination for the Improvement of Higher Education Personnel (CAPES), and in part by the Network for Analysis of Renewable Energy and Economic Development (RAERDE).

ABSTRACT The rapid rollout of industrial photovoltaic (PV) generation demands robust tools for benchmarking performance and guiding investment. Traditional efficiency techniques–non-parametric data envelopment analysis (DEA) and parametric stochastic-frontier analysis (SFA)–offer complementary strengths, yet each has recognized limitations when applied in isolation. This study proposes a hybrid framework that i) estimates multiple DEA and SFA frontiers; ii) learns their predictive patterns with Random Forest, Gradient Boosting and XGBoost models; and iii) integrates the resulting efficiency scores through an optimized weighting scheme derived from a simplex-lattice mixture design of experiments. The approach is demonstrated on a statewide dataset covering 70 immediate geographic regions of Minas Gerais, Brazil, comprising installed capacity, number of PV units, solar-radiation estimates and realized energy output. Optimization yields a parsimonious composite based primarily on the DEA and SFA models, reducing the principal-component loss metric by 18 % relative to equal weighting. Spatial analysis reveals a clear north–south efficiency gradient correlated with solar insolation; nonetheless, centrally located industrial hubs also achieve high scores, underscoring the moderating role of infrastructure. The findings highlight regions where policy support or technology upgrades could deliver the greatest marginal gains and illustrate how mixture-design optimization can reconcile diverse frontier estimates, offering a generalizable decision-analytics toolkit for renewable-energy assessment.

INDEX TERMS Energy efficiency, data envelopment analysis (DEA), stochastic frontier analysis (SFA), machine learning, mixture design of experiments (MDE), photovoltaic (PV).

I. INTRODUCTION

The decarbonization imperative crystallized in the Paris Agreement has shifted attention from incremental energy-efficiency gains to systemic transitions towards low-carbon power systems. Among commercially mature renewables, solar PVs are distinguished by their steep learning curves, modular deployment, and zero-emission operation, qualities that render them especially attractive for sun-abundant emerging economies such as Brazil. Within Brazil, the state of Minas Gerais combines expansive industrial demand with one of the continent's highest levels of global horizontal irradiance, positioning large-scale PV as a pivotal lever for simultaneous economic and environmental objectives.

Maximizing that leverage, however, demands more than expanding installed capacity; it requires a granular understanding of which installations, under which techno-economic and geographic conditions, transform solar and infrastructural inputs into electrical outputs most efficiently. Robust efficiency benchmarks enable policy makers to channel incentives, guide investors towards under-performing assets with upgrade potential, and inform grid-integration planning.

Frontier-based performance analysis has long served as the decision-analytics workhorse for such benchmarking exercises. Non-parametric DEA constructs a piece-wise linear envelope that envelops the data without a priori functional form assumptions. Yet DEA conflates random noise with managerial inefficiency, potentially penalizing units that merely experience exogenous shocks [1]. Parametric SFA remedies that limitation by decomposing the composite error term into noise and inefficiency components, but does so at the cost of imposing distributional and functional assumptions that may be mis specified in practice [2]. Recognizing these complementary strengths and weaknesses, recent studies advocate hybrid approaches that estimate multiple DEA and SFA frontiers, sometimes augmented by machine-learning surrogates, to capture a richer efficiency signal. Nonetheless, the literature still lacks a theoretically grounded mechanism for integrating those heterogeneous scores into a single, decision-relevant indicator [3].

This study addresses that gap by importing MDE from industrial statistics into the efficiency-analysis domain. By treating the weights assigned to each frontier estimate as mixture proportions and optimizing them on a multivariate loss surface, MDE yields a parsimonious yet statistically defensible composite efficiency score [4].

Using a newly assembled 2010 panel covering all 70 immediate geographic regions of Minas Gerais—including installed capacity, number of industrial PV units, GIS-derived solar-radiation indices, and realized electricity output—we demonstrate the proposed hybrid framework and reveal actionable spatial patterns in industrial PV performance.

Accordingly, the contribution of this study is primarily methodological, but with direct policy relevance. The proposed framework is intended to support regional energy planning by providing a structured efficiency indicator that can inform investment prioritization, infrastructure targeting, and the design of support actions for underperforming industrial PV regions.

In practice, regional PV performance is shaped not only by installed capacity and solar-resource availability, but also by infrastructure conditions, local industrial structure, and the ability of regions to convert technical potential into realized energy output. This creates an important planning problem: policy support directed only to already advantaged regions may reinforce existing territorial inequalities, while support directed to less efficient regions may be justified when inefficiency reflects remediable technological or infrastructural constraints.

In this sense, efficiency benchmarking can provide a decision-support tool for more balanced investment prioritization and regional energy planning. The specific contributions are four-fold:

1. Data integration: We compile the most comprehensive regional-scale dataset of industrial PV operations in Minas Gerais to date, enriching official statistics with high-resolution solar-irradiance data.
2. Multi-model benchmarking: We estimate seven DEA and SFA frontiers alongside three tree-based machine-learning models (Random Forest, Gradient Boosting Machine, XGBoost), thus capturing deterministic and stochastic efficiency facets.
3. Mixture optimization: We introduce an MDE-optimized weighting scheme that fuses the individual scores into a transparent composite metric, outperforming naive averages on a principal-component loss criterion.
4. Policy insight: We map and interpret the resulting efficiency landscape, highlighting clusters where technical assistance, infrastructure upgrades, or regulatory support could yield the greatest marginal gains.

It is important to mention that in this paper, the term PV efficiency does not refer to photovoltaic cell conversion efficiency or to plant-level operating efficiency in a narrow engineering sense. Instead, it refers to a regional industrial PV production-efficiency indicator, designed for benchmarking and planning purposes.

More specifically, the study evaluates how effectively each immediate region transforms regional PV-related inputs, such as installed capacity, number of industrial PV units, and solar-resource availability, into realized electricity output. Thus, the scope of the paper is system-level and decision-support oriented.

The remainder of the article proceeds as follows. Section II synthesizes the pertinent literature on hybrid frontier methodologies and mixture designs. Section III describes the methodological framework and data. Section IV reports the empirical results and discusses managerial and policy implications. Section V concludes with future research directions.

II. LITERATURE REVIEW

Although DEA and SFA are widely used for efficiency assessment, both approaches present inherent limitations. DEA is sensitive to outliers and does not account for statistical noise, whereas SFA depends on the specification of a functional form, which may introduce bias [5]. Furthermore, existing studies highlight that these methods face difficulties when integrating multiple performance indicators into a unified framework [6].

Recent literature has explored hybrid approaches combining DEA, SFA, and machine learning techniques to improve efficiency analysis. However, these approaches are typically applied in a fragmented manner, focusing on prediction or post-estimation analysis rather than on the systematic aggregation of efficiency scores [7].

In particular, there is a lack of structured frameworks that determine optimal weights among heterogeneous efficiency measures under a constrained optimization perspective. This gap is especially relevant in complex energy systems, where multiple efficiency estimators coexist and must be reconciled into a single decision-support index.

Therefore, this section situates the proposed DEA–SFA–Machine-Learning (ML)–Mixture framework within five streams of scholarship: (i) stand-alone frontier analyses of PV performance; (ii) parametric–non-parametric hybrids; (iii) ML-augmented frontier estimation; (iv) ensemble weighting mechanisms; and (v) mixture-design applications in efficiency studies. By synthesizing these strands, we elucidate the unresolved methodological gaps that motivate the present research.

A. FRONTIER METHODS FOR PV PRODUCTION EFFICIENCY ASSESSMENT

Early PV benchmarking studies employed DEA or SFA in isolation, exploiting their respective strengths yet inheriting their weaknesses. DEA's deterministic piece-wise linear frontier offers flexibility at the expense of conflating random shocks with inefficiency, a limitation documented in empirical PV work by [1].

Reference [8] applied a joint approach between Artificial Neural Network (ANN) and Fuzzy Data Envelopment Analysis (FDEA) to optimally select the location of new solar power plants, considering uncertainty and complexity. Reference [9] used a similar approach for the same problem, but using an integrated hierarchical combination of Principal Component Analysis (PCA) and DEA.

Reference [10] apply the Slack Based Measure (SBM) model in conjunction with Machine Learning algorithms, such as Support Vector Machine (SVM), Back Propagation Neural Network (BPNN), Decision Tree (DT), etc.

Reference [11] applied DEA to evaluate the impact of geographic conditions on photovoltaic energy panels. efficiency of geographic locations

Conversely, SFA's composite-error formulation disentangles noise from inefficiency but requires distributional and functional assumptions that may be mis specified [2]. Reference [12] applied SFA and Shephard distance function to compare solar energy companies based in China and Taiwan.

Reference [13] applied a three-stage methodology, using SFA to first define the effect of environmental factors, such as temperature, cloud cover, wind, and precipitation, that impact the inputs, and then isolate, in a second stage, the environmental effect of statistical noise. These adjusted inputs are used in DEA in the third stage.

Reference [14] applied DEA with constant returns to scale (CRS) to assess the efficiency of energy distribution companies and applied SFA as a verification method. It can be observed that, in general, the studies apply parametric and nonparametric methods independently or in subsequent stages, one serving as a complement to the other.

These methodological trade-offs have catalyzed an emergent literature on hybrid approaches. Reference [15] proposed a method called Stochastic Nonsmooth Envelopment of Data (StoNED), which is semi-parametric. StoNED aims to estimate the efficiency frontier in a single step, using convex least squares. However, because it estimates the production frontier in a single step, the method does not aim to combine different efficiency measures that represent different returns to scale, technological specifications, or even different model orientations (inputs/outputs).

B. INTEGRATING PARAMETRIC AND NON-PARAMETRIC FRONTIERS

Reference [3] provide the seminal comparative review of efforts to reconcile DEA and SFA, demonstrating that neither paradigm dominates across contexts. Subsequent contributions propose two-stage workflows in which DEA slacks inform SFA regressions or vice versa, yet they typically average the resulting scores or rely on heuristic weights, leaving the integration problem under-theorized. No study, to our knowledge, applies a statistically grounded experimental-design technique to optimize those weights - an omission addressed in the present paper.

C. MACHINE-LEARNING-AUGMENTED FRONTIER MODELLING

A second stream embeds ML within frontier analysis to capture nonlinearities and enhance predictive insight. Reference [16] combines DEA with Random Forests (RF) to rank potential Australian PV sites, while [17] integrate DEA, genetic algorithms and ML to refine industrial production frontiers. Chen et al. propose a three-stage DEA–SFA–RF pipeline for green-innovation efficiency, iteratively calibrating slacks and frontier shifts. Similar hybrid designs appear in public-health [18], finance [19] and risk-management [20] domains, underscoring ML's versatility. Nevertheless, these studies still treat multiple frontiers estimates as separate inputs to an ML predictor rather than fusing them into one interpretable composite.

D. ENSEMBLE WEIGHTING AND MIXTURE DESIGN

Ensemble learning offers a generic mechanism for combining models [21], yet off-the-shelf stacking or bagging schemes rarely honor the simplex constraint that efficiency-score weights sum to unity. References [4] and [22] detail MDE techniques precisely suited to such compositional optimization, but applications remain sparse in efficiency analysis. Outside the PV context, [23] employ majority-voting ensembles, and [24] use cross-efficiency voting to select classifiers, yet neither exploits mixture designs.

E. RESEARCH GAPS AND STUDY CONTRIBUTIONS

Three lacunae emerge from the foregoing review. First, frontier studies on PV often restrict attention to either non-parametric or parametric estimates, overlooking potential complementarities. Even hybrid models such as StoNED, while valuable, combine SFA and DEA in a sequential fashion, rather than allowing a flexible, data-driven integration of both in a single efficiency index.

Second, ML-enhanced hybrids seldom provide a transparent, policy-relevant composite efficiency score, instead shifting complexity to closed-box predictors. Third, mixture-design optimization—well suited to weight estimation under

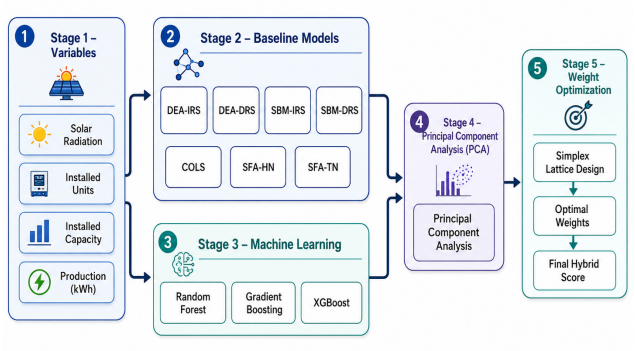

FIGURE 1. **Proposed methodology.**

convex-combination constraints—has not been leveraged to integrate DEA, SFA and ML outputs in the renewable-energy context. By introducing an MDE-optimized weighting scheme and demonstrating its utility on industrial PV data from 70 regions of Minas Gerais, this paper seeks to close these gaps.

Overall, this synthesis clarifies the methodological foci and unresolved issues that the present study seeks to overcome.

III. METHODOLOGY

To evaluate the efficiency of the intermediate regions of Minas Gerais regarding the efficiency of photovoltaic energy generation, this work applied the methodology presented in Fig. 1.

First, the raw database is filtered to industrial PV units and aggregated at the immediate-region level and the variables are selected. Second, multiple frontier models are estimated in parallel, including DEA-IRS, DEA-DRS, SBM-IRS, SBM-DRS, COLS, SFA-HN, and SFA-TN, generating a set of candidate efficiency scores. Third, these efficiency measures are learned by RF, GBM, and XGBoost models, whose predictive performance is evaluated through RMSE, MAE, and MDAE and PCA is applied to summarize the performance metrics into a single response variable. Finally, the mixture-design optimization determines the weights of the candidate efficiency models and yields the final composite efficiency index.

To provide an intuitive interpretation of the proposed framework, each component plays a distinct and complementary role. DEA identifies the relative technical efficiency of each region by comparing it with the best observed performers, without requiring a predefined functional form. SFA complements this view by estimating efficiency while explicitly accounting for statistical noise and random shocks.

The ML models are then used to learn and predict the patterns contained in the efficiency outputs generated by the frontier models, allowing their predictive performance to be assessed in a systematic way. Finally, PCA and mixture-design optimization are used to synthesize these complementary results into a single composite efficiency indicator.

A. INPUT DATA

The database used for this study was the Distributed Generation–List of Projects for the year 2010. The complete database has 285,328 observations of photovoltaic energy producers with 18 variables. The variables present in the full database are presented below.

(1) enterp_registr	Date of the latest enterprise registration update.
(2) enterp_registr	Date of the latest enterprise registration update.
(3) mun_code	Official IBGE code identifying the municipality.
(4) mun_name	Name of the municipality where the unit is located.
(5) immed_region	The immediate geographic region as defined by IBGE.
(6) cons_class	Category of electricity consumption.
(7) tariff_subgroup	Electricity tariff subgroup classification.
(8) consumer_type	Type of consumer (eg, individual, company, government).
(9) taxpayer_id	Taxpayer identification number (CPF or CNPJ).
(10) enabled_modality	Type of distributed generation modality authorized.
(11) activity_code	Main economic activity classification (CNAE).
(12) longitude	Geographic coordinate indicating east-west position.
(13) latitude	Geographic coordinate indicating north-south position.
(14) tilt_angle	The inclination angle of the installed system (in degrees).
(15) energy_prod_kwh	Estimated or actual energy production (in kWh).
(16) inst_capac_kw	Installed generation capacity of the system (in kW).
(17) econ_sector	Name of the broader economic sector (eg, agriculture, industry).
(18) econ_activity	Name of the specific economic activity according to CNAE.
(19) units_credit	Number of consumer units receiving energy credit.

We filtered the data to include only industrial consumers and aggregated the database by the immediate regions of Minas Gerais. There are 70 immediate regions in Minas Gerais. These regions are represented in Fig. 2.

In addition to methodological issues in efficiency estimation, regional PV benchmarking depends on the integration of heterogeneous data sources. In practice, administrative PV records often provide installation-level information on

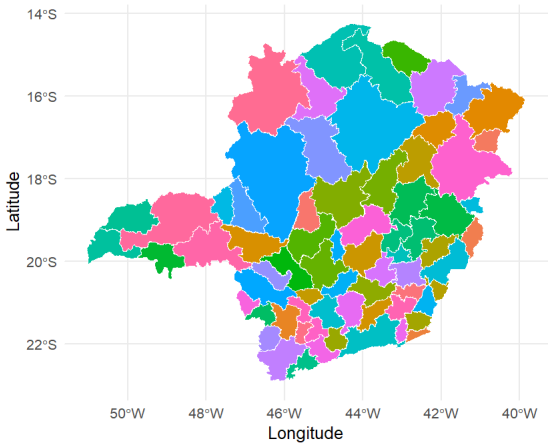

FIGURE 2. **Immediate regions of minas gerais.**

capacity, location, and consumer category, while solar-resource conditions may need to be obtained or derived from geospatial variables. This creates nontrivial analytical challenges related to data filtering, variable harmonization, spatial aggregation, and consistency between technical and geographic information.

Despite its practical importance, this data-integration stage is often treated as a background step rather than as an explicit contribution. In the present study, it plays a central role because the regional efficiency analysis requires the construction of a unified dataset combining official PV records with GIS-derived solar-radiation information.

As the literature review showed, the amount of solar radiation is an important variable for determining the efficiency of photovoltaic energy production systems. Given the longitude and latitude of each production unit, the R packages *elevatr* and *solrad* were used to estimate, respectively, the elevation to sea height and, using these data as inputs to the *OpenRadiation function* (together with the latitude and longitude), the estimate of total solar radiation in the analyzed period.

For the next step (efficiency modeling), the following were selected as *inputs*: installed capacity in kW, the number of photovoltaic energy-producing units, and the total amount of radiation in 2010. The total amount of photovoltaic energy in kWh of the immediate regions under analysis was selected as *output*. These variables constitute the *input-output* structure used in the frontier-based efficiency models.

B. BASELINE MODELS

This topic will discuss the production efficiency models used in this article, in particular, DEA and SFA. Their theoretical foundations will be addressed, in addition to the Machine Learning techniques used to predictively model the efficiency models. Then, the proposed methodology for combining models, based on design and mixture analysis, used in this work to integrate the results obtained by the DEA and SFA models trained by the machine learning algorithms presented, will be presented.

1) DATA ENVELOPMENT ANALYSIS

Discussions on efficiency analysis began with the seminal paper. Reference [25] provided definitions of technical and allocative efficiency. Based on these [26] proposed the first DEA model, called CCR (Charnes-Cooper-Rhodes).

Classical DEA models differ from each other in the imposition of economies of scale considered in the modeling. While the CCR model imposes constant returns to scale on the estimated frontier, the BCC (Banker-Charnes-Cooper) model considers variable returns to scale to define the frontier. There are also cases in which the frontier is estimated considering increasing (IRS) and decreasing (DRS) returns to scale.

Let be $x_i \in \mathbb{R}^r$ a vector of inputs and $y_j \in \mathbb{R}^m$ a vector of outputs. Equation (1) presents output-oriented DEA modeling for these models.

$$\begin{aligned} max\ \theta_0 &= \sum_{j=1}^{m} u_j y_{j0} + u_* \\ \text{S.T.:}\quad & \sum_{i=1}^{r} v_i x_{i0} = 1 \\ & -\sum_{i=1}^{r} v_i x_{ik} + \sum_{j=1}^{m} u_j y_{jk} + u_* \le 0,\ \forall k \\ & u_* = 0 : \text{CRS}; u_* \neq 0 : VRS; u_* \ge 0 : \text{IRS}; u_* \le 0 : \text{DRS} \\ & v_i; u_j \ge 0, u_* \in \text{R} \end{aligned} \tag{1}$$

For: θ_0 = efficiency of DMU_0; $v_j$eu$_j$ = weights of inputs *i, i= 1,2,..., r* and outputs *j, j= 1,2,..., m*; x_{i0}ey_{j0} = inputs *i* and outputs *j* of DMU_0. x_{ik}ey_{jk} = inputs *i* and outputs *j* of DMU_k; u_*ev_* = values of scale factors; k = 1,2,..., *n.*

Reference [27] proposed a non-radial DEA model–Slack Based Measure (SBM)–that directly considers inputs and outputs slacks $(s_r^+; s_i^-)$ without assuming proportional expansion in the output-oriented model. The mathematical formulation of the SBM model, oriented to outputs, is given by (2).

$$\begin{aligned} max\ \rho &= \frac{1 + \frac{1}{m}\sum_{j=1}^{m} \frac{s_j^+}{y_{j0}}}{1 + \frac{1}{r}\sum_{i=1}^{r} \frac{s_i^+}{x_{i0}}} \\ \text{S.T.:}\quad & x_0 = X\lambda + s^- \\ & y_0 = Y\lambda + s^+ \\ & \lambda \ge 0; s^- \ge 0; s^+ \ge 0 \end{aligned} \tag{2}$$

Considering that CRS is a particular case included in the scope of VRS, we chose to focus exclusively on IRS and DRS, prioritizing models that allow a more targeted analysis of scale effects, without analytical dispersion with models that, a priori, do not add relevant information to the objective of the study.

Therefore, DEA evaluates the relative technical efficiency of each decision-making unit (DMU) by comparing it with an empirical best-practice frontier constructed from the observed data.

Because DEA does not impose a predefined production function, it is suitable for capturing heterogeneous regional production patterns. We consider output-oriented models because the objective is to assess how effectively each region transforms installed capacity, number of units, and solar radiation into electricity generation.

We also examine alternative scale assumptions, particularly IRS and DRS, to account for possible differences in regional operating scale. In addition, SBM is included to capture non-radial inefficiency through slack information.

2) PARAMETRIC FRONTIERS

In contrast to DEA, which is a non-parametric technique, parametric techniques assume a functional form for the production function. Among these models are Corrected Ordinary Least Squares (COLS) and stochastic efficiency frontiers. Considering $x_i \in \mathbb{R}^r$ a vector of inputs and $y_j \in \mathbb{R}^m$ a vector of outputs.

It is estimated via OLS, with the output variable as the dependent and the inputs as explanatory variables. Then, the largest positive residual of the regression is identified, which represents the most efficient unit in the dataset. This value is then used to parallel transport the estimated line upwards (in the case of outputs) or downwards (in the case of inputs) so that all observations are above or below the new frontier. The technical efficiency of each DMU is then calculated as the ratio between the observed value and the estimated value at the displaced frontier. The mechanics of the COLS model are presented in (3)-(5).

$$y_i = f(x_i;\beta) + \varepsilon_i \tag{3}$$

$$\hat{\beta}_o^{COLS} = \hat{\beta}_0 + \max \varepsilon_i \tag{4}$$

The technical efficiency of DMU_i, in an output-oriented analysis, is given by:

$$\theta_i = \frac{y_i}{\hat{y}_i^{COLS}} \tag{5}$$

The SFA model is a parametric approach and is used when it is recognized that the data may be subject to random noise. The basic form of the SFA model is shown in (6).

$$y_i = f(x_i;\beta)\exp(v_i - u_i) \tag{6}$$

For: y_i = vector of outputs; x_i = vector J x1 of input variables; β = vector J x 1 of the coefficients of the vector x_i; v_i = random error with zero mean; u_i = measure of inefficiency, for $u_i \geq 0$.

In Eq. (6), the term v_i captures statistical noise and exogenous shocks, whereas u_i represents non-negative technical inefficiency. This distinction is particularly relevant in the present application because regional PV generation may be affected by measurement noise, local operating conditions, and other unobserved factors that should not be fully interpreted as inefficiency. For this reason, SFA complements DEA by providing a stochastic view of efficiency. In the proposed framework, DEA and SFA are not treated as competing methods, but as complementary sources of efficiency information.

Rather than treating these approaches as mutually exclusive, we propose here a hybrid framework in which multiple DEA and SFA models are combined based on a weighting scheme derived from MDE, in the context of a DOE. This strategy is based on the recognition that different specifications of parametric and nonparametric models produce efficiency estimates that can capture different facets of the production reality.

Therefore, the efficiency models to be applied in this work are: DEA-IRS, DEA-DRS, SBM-IRS, SBM-DRS, COLS, SFA-HN e SFA-TN. It should be noted that any efficiency measure, or any model orientation, can be applied. The resulting efficiency measures are then used as the basis for the ML-based predictive modeling stage.

C. MACHINE LEARNING

Supervised machine learning seeks to model complex relationships between predictor variables and a response variable from a labeled dataset. Among the most effective and widely used methods are algorithms based on decision trees, notably Random Forest, Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost).

1) RANDOM FOREST

Random Forest is a machine learning model that uses the *bagging method* (*bootstrap aggregating*). Consider $D = \{(x_i; y_i)\}_{i=1}^n$ a dataset with n observations, where $x_i \in \mathbb{R}^p$ are the predictors and y_i is the response variable. The algorithm consists of:

1. Generate B subsets $D^{(b)}$ via sampling with replacement D.
2. For each subset, fit a decision tree $T_b(x)$, where each division considers a random sample of $m < p$ variables.
3. The final prediction is given by:
   - Regression: $\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
   - Classification: $\hat{f}_{RF}(x) = mode\{T_b(x)\}_{b=1}^{B}$

2) GRADIENT BOOSTING MACHINE (GBM)

Gradient Boosting (Friedman, 2001) is a *boosted learning technique* where models are adjusted sequentially, each one trying to correct the errors of the previous model. The goal is to minimize a loss function $L(y; f(x))$, such as the quadratic error in the case of regression: $\hat{f} = \arg\max_f \sum_{i=1}^n L(y_i; f(x_i))$.

The model is built iteratively, so that:

1. It is initialized with a constant prediction: $f_o(x) = \arg\max_\Upsilon \sum_{i=1}^n L(y_i; \gamma)$
2. To $m = 1, 2, \ldots, M$:

The pseudo-residue is calculated

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

Adjust a tree $T_m(x)$ to the residuals $\{r_{im}\}$

The model is updated:

$$f_m(x) = f_{m-1}(x) + v\rho_m T_m(x)$$

where $v \in (0, 1]$ is the learning rate and is ρ_m the optimization coefficient of each step; $T_m(x)$ = decision tree fitted at boosting iteration m to the pseudo-residuals and; ρ_m is the step-size coefficient associated with that tree.

3) EXTREME GRADIENT BOOSTING (XGBOOST)

XGBoost is an optimized implementation of Gradient Boosting, designed for computational efficiency, explicit regularization, and scalability. It excels in data science competitions due to its accuracy and speed.

Given a prediction model $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$ where each $f_k \in \mathcal{F}$ is a decision tree, the objective is to minimize the regularized cost function (7).

$$L = \sum_{i=1}^{n} L(y_i; \hat{y}_i) + \sum_{i=1}^{K} \Omega(f_k)$$
$$For: \Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda \|w\|^2 \tag{7}$$

where T is the number of leaves in the tree f, w are the leaf weights, γ controls the complexity, and λ is the L2 regularization parameter. At step t, XGBoost expands the objective function in second-order Taylor series such that $\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)$; $g_i = \partial L\left(y_i; \hat{y}_i^{(t-1)}\right)$; $h_i = \partial^2 L\left(y_i; \hat{y}_i^{(t-1)}\right)$.

The optimal choice of the tree structure f_t is made by maximizing the gain of each split based on the values of g_i, and h_i.

4) PERFORMANCE MEASURES

When evaluating predictive models, especially in regression tasks, it is essential to quantify how close the predictions are to the actual observed values. To achieve this, several metrics are used. Among the most common are the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and the Median Absolute Error (MDAE). Each of these provides a distinct perspective on the model's performance.

The MAE measures the average magnitude of prediction errors. It is a metric that is robust to outliers and easy to interpret and is given by: $MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$.

RMSE penalizes larger errors more heavily, but is particularly sensitive to outliers and its formula is: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

The MDAE is the median of the absolute values of the forecast errors. Its formula is given by: $MDAE = median(|y_i - \hat{y}_i|)$. The predictive performance metrics obtained in this stage are subsequently summarized through PCA.

D. PRINCIPAL COMPONENT ANALYSIS

Since the performance metrics are highly correlated, Principal Component Analysis (PCA) was applied. The first principal component, which captures the largest share of the total variance, was retained and used as the response variable in the mixture design for optimizing the model weights.

Consider $X \in \mathbb{R}^{nxp}$ a data matrix with n observations and p variables, centered in relation to the mean of each column. PCA seeks to find an orthonormal basis $\{w_1, w_2, \ldots, w_p\}$ such that the first principal component $z_1 = Xw_1$ maximizes the variance of the data, as shown in (8).

$$w_1 = \arg \max_{\|w\|=1} w^T \sum w$$
$$\text{For: } \sum = \text{covariance matrix of X.} \tag{8}$$

The solution to this problem leads with the eigenvalue problem: $S w_i = \lambda_i w_i$, in which w_i form the directions of the principal components and λ_i indicate the eigenvalues that estimate the variance estimated by each component. Then $\sum_{i=1}^{p} \lambda_i$ indicates the total variance of the data explained by the i components retained in the analysis. In this work, the components that explain at least 80% of the total variance of the data will be retained.

The retained principal component is then used as the response variable in the mixture-design optimization stage.

E. WEIGHT OPTIMIZATION

DOE is a systematic statistical approach used to investigate the effects of multiple variables on one or more response variables [22].

In many situations, the factors involved in the experiment must obey a restriction: the sum of the proportions allocated to each factor must be equal to one. When this happens, we are faced with a MDE. Assuming that there are G factors and that x_k represents the proportion attributed to the k-th factor, we have the following condition presented in (9).

$$\sum_{k=1}^{G} x_k = 1 \tag{9}$$

Within MDE, the two main types of design are Simplex Lattice Design (SLD) and Simplex Centroid Design (SCD). An SLD with k components and m+ 1 equidistant levels is usually designated as SLD (k, m). The purpose of MDE in this study is not to model a physical mixture, but to optimize the weights assigned to the candidate efficiency models. Each model contributes one efficiency vector, and the corresponding weight is treated as a mixture proportion.

This makes MDE a natural framework for estimating a transparent composite efficiency index. The simplex-lattice design is used to explore the weight space systematically, allowing the estimation of a response surface by means of Ordinary Least Squares (OLS), projecting it onto sections parallel to the base of the simplex. In general, a functional

TABLE 1. presents the descriptive statistics of the selected variables.

	Mean	sd	Min	Max	Median
Installed Power (Kw)*	2.92	4.93	0.057	29.48	1.31
Production Units	57.42	94.50	2.00	562.00	26.00
Solar Radiation (W/m²)	373.99	16.31	335.60	404.02	378.01
Production (Kwh)*	1.48	2.53	0.27	151.59	6.39

* Thousands

relationship is established between the response variable and the proportions of the components, according to (10):

$$y_i = \varphi(x_1, x_2, \ldots, x_k) \tag{10}$$

where y_i is the dependent variable, $x_k, \text{for} k = 1, 2, \ldots, G$, are the explanatory variables or proportions and φ is the functional form of the model.

Since $\sum_{k=1}^{G} x_i = 1$, one degree of freedom is obtained. Then the formulation of a quadratic regression follows the functional form of (11).

$$y = \sum_{k=1}^{G} \beta_k^* x_{pk} + \sum_{k<j}^{G} \sum_{k<j}^{G} \beta_{kj}^* x_k x_j + \varepsilon \tag{11}$$

Given the estimated response surface, it is optimized using the *Constrained algorithm. Optimization by Linear Approximations* (COBYLA)–for more details about the algorithm, see [28]. The optimization function is given by (12)

$$\begin{aligned} &\max_x \ f(x) \\ &\text{S.T.:} \quad \sum_{k=1}^{G} x_k = 1 \\ &\qquad 0 \le x_k \le 1, for k = 1, 2, \ldots, G \end{aligned} \tag{12}$$

Given the optimal weights that minimize the single performance vector PCA, these weights are applied to the efficiency vectors selected to compose the weighted efficiency measure, which is given by $\theta^* = \sum_{i=1}^{G} x_i^* \theta_i$, where x_i^* and e_i represent respectively the optimal percentages and the efficiency vectors of the models selected by the optimization model.

IV. RESULTS AND DISCUSSIONS

Table 1 presents the descriptive statistics of the selected variables.

To have an overview of the productive configuration of the regions, Fig. 3 presents the map of the concentration of producers and production by the immediate region of Minas Gerais.

The immediate regions of Belo Horizonte and Divinópolis are the regions with the most photovoltaic energy producers and, therefore, the regions that produce the most photovoltaic energy in the context analyzed. However, if the average production per producing unit is analyzed, the immediate region of the São Francisco Valley and Janaúba stands out.

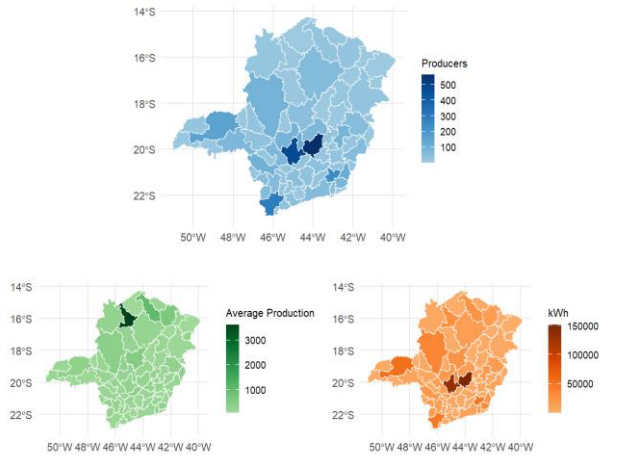

FIGURE 3. Concentration of producers, average production, and total production by immediate regions of minas gerais.

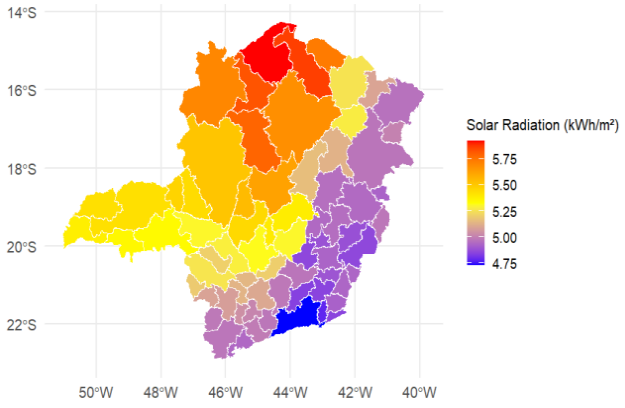

FIGURE 4. Estimated solar radiation in kWh/m^2 by immediate region of minas gerais.

TABLE 2. Descriptive statistics of efficiencies for each model applied.

Model	Min	Q1	Median	Mean	Q3	Max
IRS	0.85	0.90	0.93	0.93	0.97	1.00
DRS	0.79	0.84	0.87	0.88	0.91	1.00
SBM-IRS	0.82	0.86	0.89	0.91	0.93	1.00
SBM-DRS	0.83	0.89	0.92	0.92	0.97	1.00
COLS	0.95	0.97	0.97	0.97	0.98	1.00
SFA-HN	0.86	0.88	0.89	0.90	0.91	1.00
SFA-TN	0.95	0.97	0.97	0.97	0.98	1.00

There is a greater concentration of regions with higher productivity in the Northern region of Minas Gerais, which is explained by the high incidence of solar radiation in the region, while central regions of the state (Belo Horizonte and Divinópolis) stand out in the amount as shown in Fig. 4.

The descriptive statistics of the efficiency measures for each proposed model are presented in Table 2.

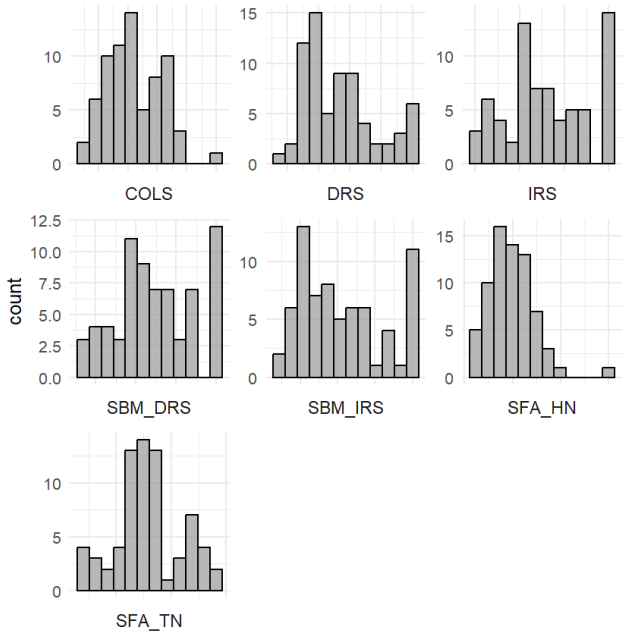

FIGURE 5. **Histograms of the results of the applied efficiency models.**

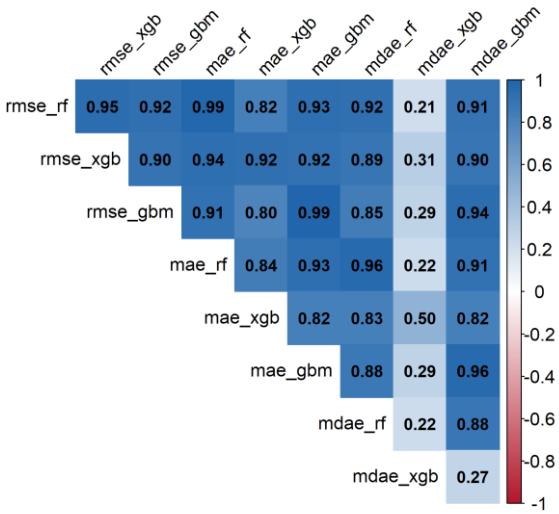

FIGURE 6. **Correlation matrix of performance measures.**

TABLE 3. Descriptive statistics of the applied ML models (cross-validation).

Variable	Mean	Sd	Min	Max
rmse_rf	0.0067	0.0018	0.0021	0.0128
rmse_xgb	0.0186	0.0014	0.0153	0.0224
rmse_gbm	0.0140	0.0040	0.0044	0.0265
mae_rf	0.0050	0.0014	0.0014	0.0092
mae_xgb	0.0149	0.0008	0.0133	0.0177
mae_gbm	0.0103	0.0031	0.0024	0.0191
mdae_rf	0.0039	0.0012	0.0009	0.0069
mdae_xgb	0.0121	0.0008	0.0105	0.0153
mdae_gbm	0.0074	0.0025	0.0013	0.0147

Fig. 5 presents the histograms of the results of the applied efficiency models (DEA-DRS, DEA-IRS, SBM-DRS, SBM-IRS, COLS, SFA-HN, and SFA-TN).

After estimating the efficiency models, each model was adjusted using Random Forest, GBM, and XGBoost, separating 80% of the database as a training set and the remaining 20% as a test set, applying *cross-validation* with $k = 5$. Table 3 presents the descriptive statistics of the performance metrics per applied model.

The aim is to assign weights to the efficiency models that jointly minimize the presented performance metrics. To assess the possibility of applying dimensionality reduction techniques, the correlation matrix of these metrics was extracted, presented in Fig. 6.

It is noted that there is a significant number of correlations above 0.30, indicating that the application of dimensionality reduction techniques may be appropriate. Bartlett's sphericity test presents a p-value of 0.000, rejecting the null hypothesis that the correlation matrix is an identity matrix.

To confirm the suitability of dimensionality reduction, the Kaiser-Meyer-Olkin (KMO) statistic was extracted, which indicated a value of 0.81, confirming the suitability of applying PCA. The first component explains 81.94% of the total variance and was therefore retained as the response variable to be optimized through the model weights.

The selected mixture design was the SLD (7, 5), adding 7 axial points and 1 central point to the experimental design so that the Response Surface model can detect curvature in the estimated quadratic regression surface. It is worth noting that the Stepwise algorithm was applied for model selection, which is presented in Table 4.

The model as a whole is significant at the 1% significance level. After estimating the Response Surface model, the COBYLA algorithm was applied, according to the model of Eq. (12). Table 5 presents the efficiency models and the correspondence concerning the symbols of the variables used in the regression model, together with the optimal values according to the optimization model.

The only models selected to compose the weighted efficiency were DEA-IRS and SBM-DRS and SFA-HN. Thus, the weighted efficiency is given by $\theta^* = 0.646\theta_{IRS} + 0.106\theta_{DRS} + 0.249\theta_{SBM-DRS} + 0.04\theta_{SFA-HN}$.

Fig. 7 shows the map of Minas Gerais subdivided into immediate regions, indicating the weighted efficiency scale calculated by the methodology proposed by this work.

As expected, the efficiency shows an increasing gradient from the south to the north of Minas Gerais, following the increase in solar radiation. This factor encourages greater investment in photovoltaic production in the region, which is reflected in the higher efficiencies observed on the map [29].

The immediate regions with the lowest efficiency in the production of photovoltaic energy for the industrial sector are Carangola, São João del Rei, and Três Corações, with a

TABLE 4. Mixture regression.

	Estimate	Std.Error	Statistic	p-value
x1	-6.9663	0.1323	-52.642	0.000***
x2	-0.7311	0.1323	-5.525	0.000***
x3	-5.0206	0.1323	-37.938	0.000***
x4	-7.6647	0.1323	-57.919	0.000***
x5	2.9424	0.1323	22.234	0.000***
x6	2.7951	0.1323	21.122	0.000***
x7	2.7972	0.1323	21,137	0.000***
x1:x2	6.0128	0.4811	12.498	0.000***
x1:x3	2.1523	0.4811	4.474	0.000***
x1:x4	1.0821	0.4811	2.249	0.025*
x1:x5	4.1183	0.4811	8.560	0.000***
x1:x6	9.6956	0.4811	20.153	0.000***
x1:x7	6.2001	0.4811	12.887	0.000***
x2:x3	5.5373	0.4811	11.510	0.000***
x2:x4	7.2416	0.4811	15.052	0.000***
x2:x5	1.2611	0.4811	2.621	0.01**
x2:x6	8.5380	0.4811	17.747	0.000***
x2:x7	3.2592	0.4811	6.774	0.000***
x3:x4	2.3692	0.4811	4.925	0.000***
x3:x5	3.3221	0.4811	6.905	0.000***
x3:x6	11.2107	0.4811	23.302	0.000***
x3:x7	3.4705	0.4811	7.214	0.000***
x4:x5	4.3019	0.4811	8.942	0.000***
x4:x6	10.2513	0.4811	21.308	0.000***
x4:x7	5.9894	0.4811	12.449	0.000***
x5:x6	10.0421	0.4811	20.873	0.000***
x5:x7	1.8592	0.4811	3.864	0.000***
x6:x7	10.3319	0.4811	21.476	0.000***

R^2 = 92.08% | R^2-Adj. = 89.07% | F-Statistic: 2118.38 | p-value: 0.000

TABLE 5. Efficiency models and optimal weight values.

Model	Symbol	Optimal Value
IRS	x_1	0.646
DRS	x_2	0.106
SBM-IRS	x_3	0.000
SBM-DRS	x_4	0.249
COLS	x_5	0.000
SFA-HN	x_6	0.040
SFA-TN	x_7	0.000

weighted efficiency score of 0.85, while the most efficient region is Januária.

Notable exceptions among the immediate regions that demonstrate high performance, despite being located in the central part of the state, include Belo Horizonte (the immediate region where the state capital is located) and Divinópolis. These regions benefit from a relatively well-consolidated industrial capital structure.

TABLE 6. Benchmarking analysis.

RMSE							
	θ_*	θ_M	θ_{DEA}	θ_{SFA}	θ_{RF}	θ_{XGB}	θ_{GBM}
IRS	0.008	0.024	0.028	0.056	0.040	0.041	0.042
DRS	0.065	0.065	0.044	0.098	0.081	0.072	0.082
SBM_IRS	0.037	0.043	0.020	0.080	0.061	0.056	0.062
SBM_DRS	0.005	0.029	0.024	0.063	0.046	0.045	0.048
COLS	0.064	0.045	0.075	0.014	0.026	0.039	0.026
SFA-HN	0.055	0.035	0.064	0.012	0.020	0.027	0.020
SFA-TN	0.067	0.047	0.078	0.012	0.028	0.040	0.027
MAE							
	θ_*	θ_M	θ_{DEA}	θ_{SFA}	θ_{RF}	θ_{XGB}	θ_{GBM}
IRS	0.006	0.019	0.020	0.048	0.035	0.035	0.037
DRS	0.044	0.055	0.029	0.088	0.072	0.065	0.074
SBM_IRS	0.021	0.036	0.011	0.070	0.054	0.049	0.055
SBM_DRS	0.003	0.022	0.015	0.053	0.039	0.038	0.041
COLS	0.053	0.039	0.065	0.011	0.024	0.038	0.024
SFA-HN	0.045	0.030	0.056	0.011	0.016	0.021	0.016
SFA-TN	0.057	0.042	0.069	0.010	0.025	0.037	0.025
MdAE							
	θ_*	θ_M	θ_{DEA}	θ_{SFA}	θ_{RF}	θ_{XGB}	θ_{GBM}
IRS	0.004	0.017	0.014	0.044	0.032	0.030	0.035
DRS	0.020	0.054	0.014	0.099	0.077	0.061	0.073
SBM_IRS	0.009	0.032	0.005	0.063	0.049	0.050	0.051
SBM_DRS	0.003	0.017	0.006	0.051	0.035	0.034	0.038
COLS	0.051	0.042	0.069	0.009	0.024	0.038	0.025
SFA-HN	0.042	0.028	0.054	0.010	0.012	0.017	0.013
SFA-TN	0.052	0.044	0.071	0.011	0.025	0.040	0.027

For: θ_* = efficiency vector obtained from the proposed method; θ_M = efficiency vector from the simple average of all model scores; θ_{DEA} = efficiency vector obtained from the DEA model only; θ_{SFA} = mean efficiency vector from SFA models; θ_{RF} = efficiency vector obtained from Random Forest; θ_{XGB} = efficiency vector obtained from XGBoost; θ_{GBM} = efficiency vector obtained from GBM.

Also worth highlighting is the immediate region of Uberlândia, in the intermediate region of Triângulo Mineiro, which is the only immediate region in this intermediate area classified as having high efficiency. The southeastern regions of the state also show positive results, driven by investments in alternative energy sources.

From a policy perspective, the results suggest that efficiency differences across regions should not be interpreted only as a reflection of solar-resource endowment. Although northern regions tend to benefit from more favorable irradiation conditions, some central industrial hubs also perform well due to stronger infrastructure and production capacity. Therefore, policy design should distinguish between structural geographic advantage and remediable

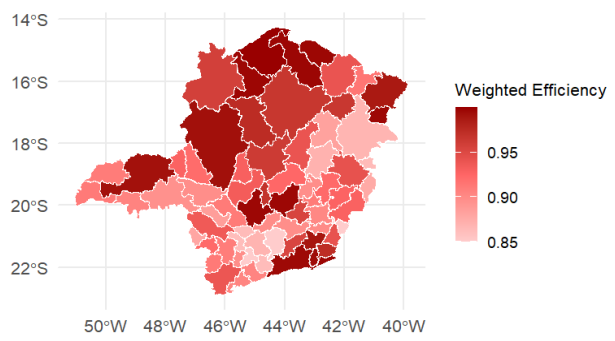

FIGURE 7. Immediate Regions with calculated weighted efficiency scale.

operational or infrastructural constraints. In this context, the proposed framework can support more targeted interventions, such as technical assistance, grid improvements, or technology-upgrade incentives, particularly in regions where performance gaps are not solely explained by natural conditions.

Public policy evaluators should be careful when directing resources to regions classified as medium-high and high-efficiency, since there is a risk of directing investments to areas that already have a competitive advantage conferred by abundant solar radiation.

Public policies focused on less favored regions could include incentives for investment in more efficient photovoltaic technologies, technical training programs for operators, and improvement of electrical infrastructure to reduce losses in the system, in addition to the implementation of hybrid systems that combine solar with other renewable sources.

This approach would make it possible to reduce regional disparities and promote more balanced energy development, avoiding the perpetuation of structural inequalities that, if ignored, can worsen the geographic concentration of investments and opportunities.

V. BENCHMARKING AGAINST ALTERNATIVES APPROACHES

To assess whether the proposed framework provides a meaningful improvement over simpler alternatives, a benchmarking analysis was carried out against representative baseline approaches. Como não há um vetor de eficiência que possa ser considerado com um vetor verdadeiro de eficiência, comparou-se os modelos em relação a eles mesmos, excluindo-se eles próprios.

Specifically, we compared the proposed mixture framework with: (i) a DEA-only benchmark based on standalone DEA specifications; (ii) a SFA-only benchmark based on standalone stochastic frontier models; (iii) a ML-only predictive baseline using RF, GBM, and XGBoost; and (iv) a simpler hybrid benchmark based on non-optimized score aggregation.

For: $\boldsymbol{\theta_*}$= efficiency vector obtained from the proposed method; $\boldsymbol{\theta_M}$= efficiency vector from the simple average of all model scores; $\boldsymbol{\theta_{DEA}}$= efficiency vector obtained from the DEA model only; $\boldsymbol{\theta_{SFA}}$ = mean efficiency vector from SFA models; $\boldsymbol{\theta_{RF}}$ = efficiency vector obtained from Random Forest; $\boldsymbol{\theta_{XGB}}$ = efficiency vector obtained from XGBoost; $\boldsymbol{\theta_{GBM}}$ = efficiency vector obtained from GBM.

All methods were evaluated under the same data partitioning, cross-validation procedure, and performance criteria. The results show whether the proposed framework improves upon isolated modeling families by combining complementary efficiency signals into a single optimized composite indicator.

The benchmarking results indicate that the proposed framework outperforms the standalone alternatives in the aggregated evaluation criterion. While DEA-only and SFA-only benchmarks capture relevant but partial views of efficiency, and ML-only provides useful predictive structure, the proposed method benefits from combining these complementary perspectives through an optimized weighting mechanism.

The proposed weighted efficiency index strongly reflects the IRS and SBM-DRS structures, as expected from the estimated weights. This is evidenced by the significantly lower approximation errors for these models. Conversely, models not included in the aggregation (e.g., SBM-IRS and SFA-TN) exhibit higher deviations, highlighting the selective nature of the proposed index.

Machine learning models outperform the proposed index in several cases due to their supervised nature. However, the proposed approach provides a structured and interpretable aggregation mechanism, avoiding closed-box approximators.

VI. CONCLUSION

This study aimed to analyze the efficiency of industrial photovoltaic solar energy generation in the immediate region of Minas Gerais, combining parametric and nonparametric approaches with Machine Learning techniques and optimization via MDE. The results demonstrated that efficiency is strongly correlated with solar radiation, with the northern regions of the state presenting the best performance due to favorable climatic conditions. The proposed methodology integrates multiple efficiency models and assigns optimal weights through optimization in the context of the Response Surface.

The application of algorithms such as Random Forest, GBM, and XGBoost allowed robust predictive modeling, while PCA and MDE ensured a balanced combination of models. The results highlight the importance of public policies and investments directed to less efficient regions, as well as the need to consider local factors, such as solar irradiation and industrial infrastructure, in energy planning.

This work contributes to the literature by proposing an innovative hybrid approach that can be replicated in other geographic contexts or energy sectors. Future research could explore the inclusion of additional variables (such as

operating costs and equipment age) and update the data to reflect recent technological advances, further expanding the applicability of the results.

Although the study demonstrates a strong correlation between solar radiation and efficiency, limitations such as the absence of recent data (post-2010) and economic variables (e.g., maintenance costs) can be explored in future work.

The combination of traditional efficiency methods with modern Machine Learning and statistical optimization techniques offers a promising path for the evaluation and continuous improvement of solar energy generation, aligning with global sustainability and energy transition goals.

REFERENCES

[1] Á. de Araújo Cavalcanti, F. de Assis dos Santos Neves, G. M. de Souza Azevedo, and A. T. de Almeida Filho, "Performance evaluation of micro- and minidistributed photovoltaic systems using data envelopment analysis," *IEEE J. Photovolt.*, vol. 9, no. 6, pp. 1806–1814, Nov. 2019, doi: 10.1109/JPHOTOV.2019.2930053.

[2] M. G. Tsionas, "Combining data envelopment analysis and stochastic frontiers via a LASSO prior," *Eur. J. Oper. Res.*, vol. 304, no. 3, pp. 1158–1166, Feb. 2023, doi: 10.1016/j.ejor.2022.04.029.

[3] C. F. Parmeter and V. Zelenyuk, "Combining the virtues of stochastic frontier and data envelopment analysis," *Oper. Res.*, vol. 67, no. 6, pp. 1628–1658, Nov. 2019, doi: 10.1287/opre.2018.1831.

[4] D. C. Montgomery, *Design and Analysis of Experiments*, 9th ed., New York, NY, USA: Wiley, 2017.

[5] S. Hanoum, M. Shubbak, D. S. Ardiantono, and J. Korpysa, "Data envelopment analysis applications for energy efficiency in power generation: A comprehensive review with bibliometric and content analysis," *Energy Convers. Manage., X*, vol. 30, May 2026, Art. no. 101590, doi: 10.1016/j.ecmx.2026.101590.

[6] F. Esteves, J. C. Cardoso, S. Leitão, and E. J. S. Pires, "Hybrid deterministic and machine learning approach for smart energy audits in wastewater treatment plants," *Discover Sustainability*, vol. 7, 2026, Art. no. 429, doi: 10.1007/s43621-026-02836-3.

[7] M. Zadmirzaei, M. Cozzi, S. Romano, A. Amirteimoori, and M. Viccaro, "Data envelopment analysis in forestry over three decades: An in-depth review of models and emerging directions," *Environ., Develop. Sustainability*, 2026, doi: 10.1007/s10668-026-07610-z.

[8] A. Azadeh, M. Sheikhalishahi, and S. M. Asadzadeh, "A flexible neural network-fuzzy data envelopment analysis approach for location optimization of solar plants with uncertainty and complexity," *Renew. Energy*, vol. 36, no. 12, pp. 3394–3401, Dec. 2011, doi: 10.1016/j.renene.2011.05.018.

[9] A. Azadeh, S. F. Ghaderi, and A. Maghsoudi, "Location optimization of solar plants by an integrated hierarchical DEA PCA approach," *Energy Policy*, vol. 36, no. 10, pp. 3993–4004, Oct. 2008, doi: 10.1016/j.enpol.2008.05.034.

[10] H. Niu, Z. Zhang, and M. Luo, "Evaluation and prediction of low-carbon economic efficiency in China, Japan and South Korea: Based on DEA and machine learning," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, p. 12709, Oct. 2022, doi: 10.3390/ijerph191912709.

[11] M. Demirkiran and A. Karakaya, "Efficiency analysis of photovoltaic systems installed in different geographical locations," *Open Chem.*, vol. 20, no. 1, pp. 748–758, Aug. 2022, doi: 10.1515/chem-2022-0190.

[12] J.-S. Wu, "Applying stochastic frontier analysis to measure the operating efficiency of solar energy companies in China and Taiwan," *Polish J. Environ. Stud.*, vol. 29, no. 5, pp. 3385–3393, 2020, doi: 10.15244/pjoes/115209.

[13] Z. Wang, Y. Li, K. Wang, and Z. Huang, "Environment-adjusted operational performance evaluation of solar photovoltaic power plants: A three stage efficiency analysis," *Renew. Sustain. Energy Rev.*, vol. 76, pp. 1153–1162, Sep. 2017, doi: 10.1016/j.rser.2017.03.119.

[14] C. von Hirschhausenand A. Cullmann, "Efficiency analysis of german electricity distribution utilities: non-parametric and parametric tests," DresdenDiscussion Paper Series in Economics, Tech. Rep. 06/05, 2005. Accessed: May 24, 2026. [Online]. Available: https://hdl.handle.net/10419/22723

[15] T. Kuosmanen and M. Kortelainen, "Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints," *J. Productiv. Anal.*, vol. 38, no. 1, pp. 11–28, Aug. 2012, doi: 10.1007/s11123-010-0201-3.

[16] G. Cattani, "Combining data envelopment analysis and random forest for selecting optimal locations of solar PV plants," *Energy AI*, vol. 11, Jan. 2023, Art. no. 100222, doi: 10.1016/j.egyai.2022.100222.

[17] J. H. Cavalcanti, T. Kovacs, A. Ko, and K. Pocsarovszky, "Production system efficiency optimization through application of a hybrid artificial intelligence solution," *Int. J. Comput. Integr. Manuf.*, vol. 37, no. 6, pp. 790–807, Jun. 2024, doi: 10.1080/0951192x.2023.2257661.

[18] O. Cosgun, G. Ogcu Kaya, and C. Cosgun, "COVID-19 vaccination performance of the U.S. states: A hybrid model of DEA and ensemble machine learning methods," *Ann. Oper. Res.*, vol. 341, no. 1, pp. 699–729, Oct. 2024, doi: 10.1007/s10479-024-06008-2.

[19] H. H. Thabet, S. M. Darwish, and G. M. Ali, "Measuring the efficiency of banks using high-performance ensemble technique," *Neural Comput. Appl.*, vol. 36, no. 27, pp. 16797–16815, Sep. 2024, doi: 10.1007/s00521-024-09929-y.

[20] S. Jomthanachai, W.-P. Wong, and C.-P. Lim, "An application of data envelopment analysis and machine learning approach to risk management," *IEEE Access*, vol. 9, pp. 85978–85994, 2021, doi: 10.1109/ACCESS.2021.3087623.

[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[22] J. Lawson, *Design and Analysis of Experiments With R*. Provo, UT, USA: Provo, 2014.

[23] A. Kemiveš, L. Barjaktarović, M. Randelović, M. Čabarkapa, and D. Randelović, "Assessing the efficiency of foreign investment in a certification procedure using an ensemble machine learning model," *Mathematics*, vol. 12, no. 7, p. 1020, Mar. 2024, doi: 10.3390/math12071020.

[24] Q. An, S. Huang, Y. Han, and Y. Zhu, "Ensemble learning method for classification: Integrating data envelopment analysis with machine learning," *Comput. Operations Res.*, vol. 169, Sep. 2024, Art. no. 106739, doi: 10.1016/j.cor.2024.106739.

[25] M. J. Farrell, "The measurement of productive efficiency," *J. Royal Stat. Soc., Ser. A*, vol. 120, no. 3, pp. 253–281, May 1957.

[26] A. Charnes, W. W. Cooper, and E. Rhodes, "A Data Envelopment Analysis Approach to Evaluation of the Program Follow through Experiment in U.S. Public School Education," *Manage. Sci.*, vol. 27, no. 6, pp. 668–697, 1978.

[27] K. Tone, "Theory and methodology a slacks-based measure of efficiency in data envelopment analysis," *Eur. J. Oper. Res.*, vol. 130, no. 3, pp. 498–509, 2011. [Online]. Available: www.elsevier.com/locate/dsw

[28] A. Tröltzsch, "A sequential quadratic programming algorithm for equality-constrained optimization without derivatives," *Optim. Lett.*, vol. 10, no. 2, pp. 383–399, Feb. 2016, doi: 10.1007/s11590-014-0830-y.

[29] R. J. dos Reis and C. Tiba, *Atlas Solarimétrico de Minas Gerais, I.*, vol. 2. Brasília, Brazil: Belo Horizonte, 2016.

GUSTAVO S. LEAL received the B.Sc. degree in economics and the M.Sc. degree in administration from the Federal University of Juiz de Fora (UFJF) and the Ph.D. degree in production engineering from the Federal University of Itajubá (UNIFEI), Brazil. He completed a specialization in computational statistical methods at UFJF. His research interests include decision support systems, optimization, productive efficiency, and financial administration.

VICTOR E. M. VALÉRIO received the degree in economic sciences from São Paulo State University, Brazil, in 2011, and the M.Sc. and Ph.D. degrees in production engineering from the Federal University of Itajubá, Brazil, in 2015 and 2018, respectively. He is currently an Adjunct Professor with the Federal University of Itajubá.

PEDRO P. BALESTRASSI received the M.Sc. degree in electrical engineering from the Federal University of Itajubá (UNIFEI), Brazil, and the Ph.D. degree in industrial engineering from the Federal University of Santa Catarina. He is currently a Full Professor with UNIFEI. His work focuses on applied statistics, quality engineering, and neural networks.

LUPÉRCIO F. BESSEGATO received the B.Sc. degree in electrical engineering and the M.Sc. and Ph.D. degrees in statistics from the Federal University of Minas Gerais (UFMG), Brazil. He is currently an Associate Professor with UFJF. His research interests in statistical learning, multivariate data modeling, industry 4.0, and statistics education.

FARID MELGANI (Fellow, IEEE) received the Ph.D. degree in electronic and computer engineering from the University of Genoa, Italy. He is currently a Full Professor with the University of Trento, Italy, where he leads the Signal Processing and Recognition Laboratory. His research interests include machine learning, remote sensing, and image/signal processing.

 Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614