Probability-based Technical Efficiency Analysis through Machine Learning

Abstract

This paper presents a novel integration of Data Envelopment Analysis (DEA) and Machine Learning (ML) to improve the robustness and interpretability of efficiency assessments for decision-making units (DMUs). The contribution to Artificial Intelligence lies in the development of a classification-based framework that reformulates DEA as a probabilistic classification task. Applying neural networks, our approach estimates not only the (in)efficiency status of each DMU but also the probability of being classified as efficient, thereby enabling counterfactual-based efficiency benchmarking. We incorporate Explainable AI (XAI) techniques—specifically sensitivity analysis and counterfactual explanations—to enhance the interpretability of the model and provide actionable insights through variable importance measures. Additionally, the framework introduces a dynamic peer selection mechanism based on probability thresholds, offering adaptable and more informative benchmarking strategies. In this paper, the engineering application focuses on the food industry in Spain, where the proposed ML-DEA hybrid method is used to evaluate and improve firm-level operational efficiency. The results demonstrate significant advantages over traditional DEA models, including improved discrimination among DMUs, robustness to data noise, and enhanced decision support through tailored efficiency improvement recommendations. This work highlights the potential of ML-enhanced DEA as a practical tool for performance analysis in industrial settings.

Keywords: Data Envelopment Analysis, Machine Learning, Classification models, Neural Networks, variable importance.

1. Introduction

In recent decades, the field of efficiency analysis has witnessed significant advancements, particularly in the evaluation of firms, institutions and organizations across various sectors such as finance, healthcare, education, and manufacturing. One prominent methodology that has garnered substantial attention is Data Envelopment Analysis (DEA), initially introduced by Charnes, Cooper, and Rhodes in the late 1970s (Charnes et al., 1978). DEA offers a non-parametric approach to assess the relative efficiency of Decision Making Units (DMUs) by comparing their input-output profiles. The fundamental premise of DEA lies in its ability to evaluate the efficiency of DMUs that operate under multiple inputs and outputs, without imposing restrictive assumptions about functional forms or underlying distributions. This characteristic makes DEA particularly appealing for analyzing complex real-world systems where the relationships between inputs and outputs are likely nonlinear and unknown. Over the years, DEA has been applied to diverse domains, including banking (Berger et al., 1997), healthcare (Olesen et al., 2007), mining and mineral resources (Khademian, 2024) and environmental performance assessment (Zhou et al., 2008), among others.

However, despite its widespread adoption and commendable performance, traditional DEA approaches may encounter limitations in capturing the intricate patterns and structures characterizing involved processes and datasets (see, for example, a recent criticism in Sahil et al., 2025). One notable challenge lies in the potential for overfitting, wherein the model captures noise or idiosyncratic features in the data rather than true underlying relationships (Esteve et al., 2020). This issue is particularly pronounced in DEA when dealing with high-dimensional datasets or when the number of DMUs is relatively small compared to the number of inputs and outputs, where overfitting is mixed with the curse of dimensionality problem (Charles et al., 2019). Then, DEA can lead to inflated efficiency scores for certain DMUs, thereby distorting the assessment of relative efficiency and potentially misleading decision-makers. Moreover, traditional DEA models rely on linear programming techniques to estimate efficiency scores, which may not adequately capture nonlinear relationships or interactions among inputs and outputs. As a result, the model may overlook certain patterns in the data, leading to biased efficiency estimates. Another significant limitation of traditional DEA is its deterministic nature. Standard DEA models produce a single efficiency score for each DMU based on the observed input-output data, without accounting for the volatile, uncertain, complex, and ambiguous (VUCA) scenarios inherently observed in real-world systems. This deterministic approach fails to acknowledge the stochastic nature of many decision-making processes.

With the advent of machine learning (ML) techniques, there exists a compelling opportunity to enhance the capabilities of DEA by exploiting the computational power and flexibility offered by these methods. By integrating ML algorithms with DEA, researchers can improve the accuracy, robustness, and interpretability of efficiency assessments, thereby advancing the state-of-the-art in performance

analysis. Nowadays, machine learning algorithms complement DEA by providing advanced techniques for, *inter alia*, data preprocessing (Chen et al., 2014), variable importance measurement (Valero-Carreras et al., 2024), and the treatment of the curse of dimensionality (Esteve et al., 2023), thereby facilitating more accurate and comprehensive efficiency assessments.

In the literature, several bridges between ML and DEA have already been established. However, there exist certain gaps that we believe the novel approach introduced in this paper can address. Before mentioning these gaps, we briefly review the main contributions relating ML and DEA. As we are aware, there are two predominant streams of research in the literature that explore their integration.[1]

The first stream focuses on adapting existing ML techniques to ensure that the predictive function, typically representing a production function in our context, complies with various shape constraints, such as monotonicity or concavity, when capturing the underlying relationships between inputs and outputs. Milestones in this domain are the following: Kuosmanen and Johnson (2010) demonstrated the connection between DEA and least-squares regression, introducing Corrected Concave Nonparametric Least Squares (C^2NLS). Parmeter and Racine (2013) proposed innovative smooth constrained nonparametric frontier estimators, incorporating production theory axioms. Daouia et al. (2016) introduced a method using constrained polynomial spline smoothing for data envelopment fitting, enhancing precision and smoothness. Esteve et al. (2020) and Aparicio et al. (2021) developed Efficiency Analysis Trees (EAT), improving production frontier estimation through decision trees. Valero-Carreras et al. (2021) introduced Support Vector Frontiers (SVF), adapting Support Vector Regression for production function estimation. Olesen and Ruggiero (2022) proposed hinging hyperplanes as a nonparametric estimator for production functions. Guerrero et al. (2022) introduced Data Envelopment Analysis-based Machines (DEAM) for estimating polyhedral technologies. Valero-Carreras et al. (2022) adapted SVF to multi-output scenarios, improving efficiency measurement. Guillen et al. (2023a, 2023b, 2023c, 2024) introduced boosting techniques for efficiency estimation in different scenarios. Tsionas et al. (2023) proposed a Bayesian Artificial Neural Network approach for frontier efficiency analysis under shape constraints. And Liao et al. (2024) proposed Convex Support Vector Regression (CSVR) to improve predictive accuracy and robustness in nonparametric regression, among others.

The second stream of literature adopts a two-stage approach to directly integrate DEA with ML techniques. In the first stage, researchers apply a pre-existing DEA model, such as the output-oriented radial model, to compute efficiency scores for each observation in the sample. In the second stage, the

[1]A third line of research in the literature, unrelated to this study, employs Data Envelopment Analysis (DEA) as an alternative method to conventional ML classification techniques such as Support Vector Machines, Decision Trees, and Neural Networks. In that line, DEA is utilized to classify observations based on their features instead of measuring technical efficiency. For example, it is applied to identify individuals as carriers of a rare genetic disorder from age and several blood measurements. A recent example of this type of contributions is Jin et al. (2024).

efficiency scores obtained from DEA are treated as the response variable in a 'regression' model based on standard ML techniques (without shape constraints). The original inputs and outputs, along with potentially additional environmental variables, serve as predictor variables in the regression model. By incorporating ML techniques into the performance evaluation framework, researchers aim to develop more robust and accurate predictive models for assessing efficiency. Some of these contributions are the following: Emrouznejad and Shale (2009) explored a novel approach by combining a neural network with DEA to address the computational challenges posed by large datasets. Liu et al. (2013) compared standard DEA, three-stage DEA, and neural network approaches to measure the technical efficiency of 29 semi-conductor firms in Taiwan. Fallahpour et al. (2016) presented an integrated model for green supplier selection under a fuzzy environment, combining DEA with genetic programming to address the shortcomings of previous DEA models in supplier evaluation. Kwon et al. (2016) explored a novel method of performance measurement and prediction by integrating DEA and neural networks. The study used longitudinal data from Japanese electronics manufacturing firms to show the effectiveness of this combined approach. Aydin and Yurdakul (2020) introduced a three-staged framework utilizing Weighted Stochastic Imprecise Data Envelopment Analysis and ML algorithms to assess the performance of 142 countries against the COVID-19 pandemic. Tayal et al. (2020) presented an integrated framework for identifying sustainable manufacturing layouts using Big Data Analytics, ML, Hybrid Meta-heuristic and DEA. The paper by Nandy and Singh (2020) presented a hybrid approach utilizing DEA and Machine Learning, specifically the Random Forest (RF) algorithm, to evaluate and predict farm efficiency among paddy producers in rural eastern India. Zhu et al. (2021) proposed a novel approach that combines DEA with ML algorithms to measure and predict the efficiency of Chinese manufacturing companies. Jomthanachai et al. (2021) proposed an integrated method combining both techniques for risk management. Boubaker et al. (2023) proposed a novel method for estimating a common set of weights based on regression analysis (such as Tobit, LASSO, and Random Forest regression) for DEA to predict the performance of over 5400 Vietnamese micro, small and medium enterprises. Amirteimoori et al. (2023) introduced a novel modified Fuzzy Undesirable Non-discretionary DEA model combined with artificial intelligence algorithms to analyze environmental efficiency and predict optimal values for inefficient DMUs, focusing on CO_2 emissions in forest management systems. Lin and Lu (2024) presented a novel analytical framework utilizing inverse Data Envelopment Analysis and ML algorithms to evaluate and predict suppliers' performance in a sustainable supply chain context. Omrani et al. (2024) evaluated the efficiency of electricity distribution companies (EDCs) from 2011 to 2020 using a combination of DEA, corrected ordinary least squares (COLS), and machine learning techniques. In particular, a three-stage process involving DEA, COLS, support vector regression (SVR), fuzzy triangular numbers, and fuzzy TOPSIS methods were employed, revealing trends in EDC performance and identifying areas needing improvement.

Both streams of research have contributed valuable insights and methodologies for integrating ML with DEA. However, despite these advancements, there remain relevant unaddressed questions that offer opportunities for innovation and further improvement. Particularly, the adoption of a probabilistic framework not only when classifying observations as efficient or inefficient, but also when improving their probability of being efficient by changing their production processes according to a given technical inefficiency measure. In this regard, our approach introduces novel functionalities and complementary methodologies to traditional DEA-based analysis by taking advantage of the favorable properties exhibited by classification models in ML. Next, we outline the key contributions of this study, highlighting its methodological innovations, interpretative advantages, and practical implications for efficiency measurement:

- *Methodological Innovation*: As we are aware, for the first time in the literature, we propose a classification-based machine learning approach in the second stage of a DEA-ML hybrid framework, moving beyond the conventional regression-based techniques. In the first stage, we rely on the concept of Pareto-dominance to distinguish between two classes of DMUs: efficient and inefficient. A DMU is Pareto efficient if and only if it is impossible to improve any input or output without worsening some other input or output. This initial classification is effectively performed with the standard additive DEA model. However, in contrast to regression-based methods that, in the first stage, require for each DMU a precise numerical efficiency score—i.e., a real valued dependent variable, which is subsequently regressed against the set of input-output and ancillary variables, our approach circumvents the challenges and potential inaccuracies associated with predicting exact numerical outcomes. By adopting a robust binary classification framework as the response variable (efficient vs. inefficient), we eliminate the risk of propagating errors inherent in continuous value predictions, thereby providing a clearer and more reliable distinction between the two classes. In the second stage, we apply classification models to predict the probability of being classified as efficient using the input and output variables.

- *Inferential Power with Probability Estimation*: One of the key advantages of our classification-based approach is that it enables the estimation of the probability of a DMU being classified as efficient. In our novel approach this probability constitutes the real valued response variable—i.e., the output of a classification method based on machine learning, that enables efficiency measurement. What is utterly relevant, and unlike traditional DEA models that mainly serve as descriptive tools, our framework incorporates the probabilistic perspective, allowing researchers and practitioners to infer efficiency status based on statistical learning principles. This conceptual shift aligns efficiency measurement with modern inferential analytics, bridging the gap between efficiency measurement and probability-based decision-making.

- *Reinterpreting DEA as a Classification Problem*: Our approach allows reframing traditional DEA as a classification problem. First, the DEA technology differentiates the variable space between two regions: technically feasible and technically infeasible production processes. Second, DEA technical efficiency measures allow classifying DMUs into the efficient and inefficient classes through the identification of the Pareto-efficient production frontier, which can be interpreted as the separating surface between the feasible and unfeasible categories. Under this reinterpretation, technical efficiency measures can be seen as quantifying the minimum required input and/or output modifications necessary for an inefficient unit to transition from the inefficient class to the efficient class, coinciding with the separating Pareto-efficient frontier. In contrast, the new method estimates the probability value at which one specific input-output combination belongs to the efficient boundary of the feasible region. As a result, for any pre-defined level of probability, our method allows classifying DMUs into efficient or inefficient classes. Additionally, under the new paradigm, technical inefficiency measures can be interpreted as distances to the efficient boundary at a pre-defined probability threshold.

- *Algorithm-Agnostic Approach for Robust Efficiency Assessments*: A key advantage of our framework is its flexibility in algorithm selection. Unlike conventional DEA-ML models that rely on a specific regression technique, our method is not tied to a particular classification algorithm. This flexibility might eventually allow us to experiment with a wider range of ML models—including decision trees, Support Vector Machines, Neural Networks, and ensemble methods—ensuring that the results remain robust and consistent across multiple techniques. Nevertheless, for the sake of simplicity, in this seminal paper we focus on Neural Networks.

- *XAI and Counterfactuals*: Another major contribution of our study is the integration of Explainable Artificial Intelligence (XAI), specifically counterfactual methods, into efficiency analysis. Instead of relying solely on conventional efficiency scores, we define technical inefficiency for an inefficient DMU as the minimum changes required in inputs and/or outputs (depending on the selected orientation) to transition with some probability from the inefficient status to the efficient one. These counterfactual-based adjustments, which can be set at different predefined thresholds of probability of being efficient, offer an intuitive and interpretable way to assess inefficiency. For illustrative purposes, and given its increasing popularity in efficiency studies, we choose the Directional Distance Function (Chambers et al., 1998) as the reference measure allowing for efficiency improvements through joint input reductions and output increments.

- *Benchmarking with Variable Importance and Directed Projections*: Our methodology also introduces a novel benchmarking approach by offering a ranking of importance of the inputs and outputs, identified through machine learning models. As highlighted in the literature (e.g., Banker and Morey, 1986; Thanassoulis et al., 2015), understanding the relative importance of variables in

efficiency assessments is crucial for strategic decision-making. We propose using this information to assign data-driven weights to inputs and outputs, guiding the projection of inefficient DMUs towards more meaningful and customized improvement paths (directional vectors). This is a significant departure from traditional DEA projections grounded on the Directional Distance Function, which often rely on subjective or arbitrary directional vectors (Wang et al., 2019).

- *Target Setting Through Counterfactual Benchmarking*: The benchmarking framework we propose is further enhanced by incorporating probabilistic efficiency thresholds and applying counterfactual analysis to determine the minimum necessary changes in inputs and outputs that would allow a DMU to be reclassified as efficient. This technique generates concrete improvement targets, calculated as the closest efficient peer at a given probability threshold (see below), and also allows practitioners to prioritize adjustments based on their impact on efficiency classification.

- *Ranking DMUs and Confidence Thresholds*: Expanding on previous works (e.g., Sexton, 1986; Thanassoulis et al., 2008), we propose a novel ranking system for DMUs based on their probabilistic efficiency scores. By exploiting the information provided by directional projections onto the separating surface of the two classes (efficient vs inefficient), we can rank units according to their likelihood of being classified as efficient, providing an interesting evaluation framework compared to traditional DEA-based ranking methods.

- *Proximity-Based Peer Identification*: Finally, our approach facilitates a more refined peer selection process by identifying, for each DMU and at every efficiency probability threshold, the closest efficient benchmark unit. This selection is performed using proximity-based metrics, such as Euclidean distance, ensuring that benchmark comparisons are contextually relevant and practically achievable.

The paper is structured as follows: In Section 2, we provide background information on Data Envelopment Analysis (DEA) and the machine learning techniques we employ, (Artificial) Neural Networks (NN). Section 3 introduces our novel approach, which integrates DEA with the classification technique, aiming to enhance efficiency assessment for DMUs. We demonstrate the practical implications of this integration and its implications for decision-making through an empirical example based on SABI (Iberian Balance Sheet Analysis System) in Section 4. Section 5 concludes and identifies further research lines.

2. Background

This background section provides a concise overview of DEA, the classifying ML technique of choice, corresponding to Neural Networks, and certain Explainable Artificial Intelligence techniques.

2.1. Data Envelopment Analysis, DEA

As aforementioned, DEA is a non-parametric method widely used for evaluating the relative efficiency of a set of observations or DMUs. DEA offers a powerful framework for assessing the efficiency of DMUs transforming multiple inputs into multiple outputs. From a technological perspective, DEA offers great flexibility when modeling the production technology, e.g., convex or non-convex, strong or weak disposability of inputs and outputs, constant or variable returns to scale, etc. (Orea and Zofío, 2019). In what follows we rely on the canonical model characterized by convexity, strong disposability and variable returns to scale. The former feature is particularly suited for analyzing real-world production processes where returns to scale may vary across different units.

The process of evaluating the technical efficiency of n observations requires comparing the performance of this dataset of DMUs (D) in terms of the m inputs that they use, $\mathbf{x}_j = \left(x_{1j},\ldots,x_{mj}\right) \in R_+^m$, such as labor, capital, and other resources, to generate s outputs $\mathbf{y}_j = \left(y_{1j},\ldots,y_{sj}\right) \in R_+^s$, like goods or services. In this notation, input and output vectors for a specific observation j are presented in bold typeface. In a conceptual framework, the term 'technology' encompasses all feasible input-output combinations. This concept is typically represented as:

$$T = \left\{(\mathbf{x},\mathbf{y}) \in R_+^{m+s} : \mathbf{x} \text{ can produce } \mathbf{y}\right\}. \tag{1}$$

Based on the n observed input-output combinations, DEA empirically approximates T through the following mathematical expression, Banker et al. (1984):

$$T_{DEA} = \left\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}_+^{m+s} : y_r \le \sum_{j=1}^{n} \lambda_j y_{rj}, \forall r, x_i \ge \sum_{j=1}^{n} \lambda_j x_{ij}, \forall i, \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \ge 0, \forall j\right\}. \tag{2}$$

The DEA technology differentiates the input-output variable space (i.e., the non-negative orthant $\mathbb{R}_+^{m+s}$) in two regions: technically feasible and technically infeasible production processes. Within the feasible region represented by T_{DEA}, numerous technical efficiency measures are available to calculate the performance of observations—for a general definition of these measures see Pastor et al. (2012). In this study we choose one of the most prominent measures, the Directional Distance Function (DDF) proposed by Chambers et al. (1998). For a specific DMU $\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right)$, and associated directional vector $\left(\boldsymbol{g}_o^x, \boldsymbol{g}_o^y\right)$ specifying the direction for its projection onto the efficient frontier of T_{DEA}, its technical efficiency corresponds to the distance between the observation and the frontier, which is be calculated through the following program:

$$
\begin{aligned}
\vec{D}\left(\boldsymbol{x}_o, \boldsymbol{y}_o; \boldsymbol{g}_o^x, \boldsymbol{g}_o^y\right) = \quad & \max \quad \beta & & (3.0) \\
& s.t. \quad \sum_{j=1}^{n} \lambda_{jo} x_{ij} \leq x_{io} - \beta g_{i0}^x, \quad i = 1,...,m & & (3.1) \\
& \sum_{j=1}^{n} \lambda_{jo} y_{rj} \geq y_{ro} + \beta g_{r0}^y, \quad r = 1,...,s & & (3.2) \\
& \sum_{j=1}^{n} \lambda_{jo} = 1, & & (3.3) \\
& \lambda_{jo} \geq 0, \quad j = 1,...,n & & (3.4)
\end{aligned}
\tag{3}
$$

A DMU with an efficiency score strictly greater than zero, i.e., $\vec{D}\left(\boldsymbol{x}_o, \boldsymbol{y}_o; \boldsymbol{g}_o^x, \boldsymbol{g}_o^y\right) = \beta^* > 0$ (with * indicating optimality) is technically inefficient relative to the reference technology T_{DEA}, with a bigger value indicating larger technical inefficiency. On the contrary, a DMU with a zero score, i.e., $\vec{D}\left(\boldsymbol{x}_o, \boldsymbol{y}_o; \boldsymbol{g}_o^x, \boldsymbol{g}_o^y\right) = \beta^* = 0$, is considered efficient, signaling that it operates on the efficient frontier of T_{DEA}—i.e., the boundary separating the feasible and infeasible regions. That is, it is impossible to simultaneously reduce inputs and increase outputs given the technology.

2.2. Neural Networks, NNs

Here we outline the fundamentals of NNs as our machine learning technique of choice to undertake classification tasks, highlighting their versatility, theoretical foundations, and practical implications. Furthermore, throughout this section, we will offer brief suggestions on how this technique could be adapted to measure the probability of efficiency for the analyzed DMUs. Indeed, this specific adaptation constitutes the core focus of Section 3 of this paper.

NNs represent a cornerstone in the field of machine learning, recognized for their ability to learn complex patterns and relationships from data (LeCun et al., 2015; Goodfellow et al., 2016). NNs are inspired by the structure and function of the human brain, comprising interconnected layers of artificial neurons or nodes. These neurons process learning data (which corresponds here to the inputs and outputs of the DMUs characterizing the technology (1)) through nonlinear transformations to learn complex patterns and relationships and predict the probability of belonging to two possible classes in the classical classification problem (being efficient vs being inefficient in our production context). The core principle underlying NNs is the process of propagation, where input data is sequentially passed through multiple layers of neurons, each layer applying a set of weights and activation functions to produce a response variable or output. Through an iterative process known as backpropagation, NNs adjust the weights of connections between neurons based on the error between predicted and actual outputs, thereby minimizing a certain loss function and improving predictive accuracy. In this sense, activation functions play a crucial role in NNs by introducing non-linearity into the model, enabling it

to capture complex relationships within the data. Common activation functions include the sigmoidal function, hyperbolic tangent (tanh) function, and rectified linear unit (ReLU) function. Each activation function introduces different properties to the model, influencing its ability to learn and generalize from data.

In a binary classification problem as the one we propose in this paper to identify efficient and inefficient observations, the neural network is designed to distinguish between two possible classes: a positive class: $y=+1$ (efficient DMUs) and a negative class: $y=-1$ (inefficient DMUs). The network processes an input $\boldsymbol{x}$ through multiple layers of neurons, applying weighted transformations and nonlinear activation functions. The hidden layers compute intermediate representations:

$$\boldsymbol{h}^{(l)} = f\left(\boldsymbol{W}^{l}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}\right), \tag{4}$$

where $\boldsymbol{W}^{l}$ and $\boldsymbol{b}^{(l)}$ are the weight matrix and bias vector at layer l, and $f(\cdot)$ is a nonlinear activation function. At the final layer, a single neuron outputs a logit z, which is mapped to a probability using the sigmoid function (the probability of being efficient):

$$P(y=+1|\,\boldsymbol{x}) = \sigma(z) = \frac{1}{1+e^{-z}}. \tag{5}$$

This probability represents the likelihood that the given input (in this case each DMU) belongs to the positive class $(y=+1)$. Since there are only two possible outcomes, the probability of the negative class $(y=-1)$ is simply: $P(y=-1|\,\boldsymbol{x}) = 1-P(y=+1|\,\boldsymbol{x})$.

The performance of NNs hinges on the selection of hyperparameters such as the number of layers, the number of neurons per layer, learning rate, and regularization parameters. Hyperparameter tuning is essential to optimize model performance and prevent issues like overfitting or underfitting.

An illustrative example of the configuration of a neural network in the context of a binary classification problem, with two predictor variables, would consist of two neurons in the input layer, reflecting the number of variables involved in the model. In our production context. these two variables would be the input and output of the production process (I1 and I2). In the response layer, a single neuron would be located to assign the corresponding class to each observation (O1): efficient or inefficient with some probability. Between these layers lies, in this case, one hidden layer, composed of a pre-defined number of neurons, three in this case (H1, H2, H3). Figure 1 depicts the structure of this neural network with a configuration of 2-3-1 and its corresponding weights.

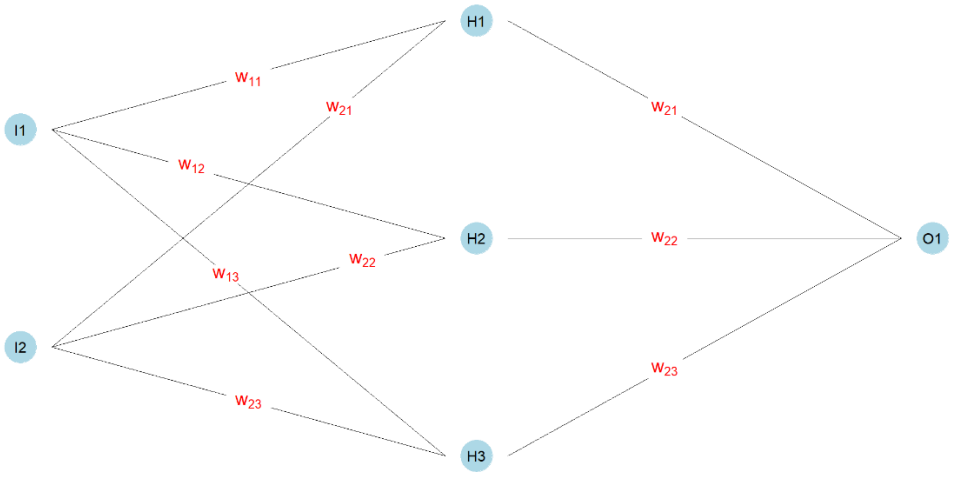

Figure 1. An example of a very simple artificial Neural Network.

Figure 2 illustrates the nonlinear decision boundary (separating surface) generated by the neural network, which partitions the feature space into two distinct regions. One region is associated with class $y = +1$ (i.e., efficient class), while the other corresponds to class $y = -1$ (i.e., inefficient class). This separating surface emerges as a result of the network's learned transformations, effectively capturing complex patterns in the data that a linear classifier would fail to model.

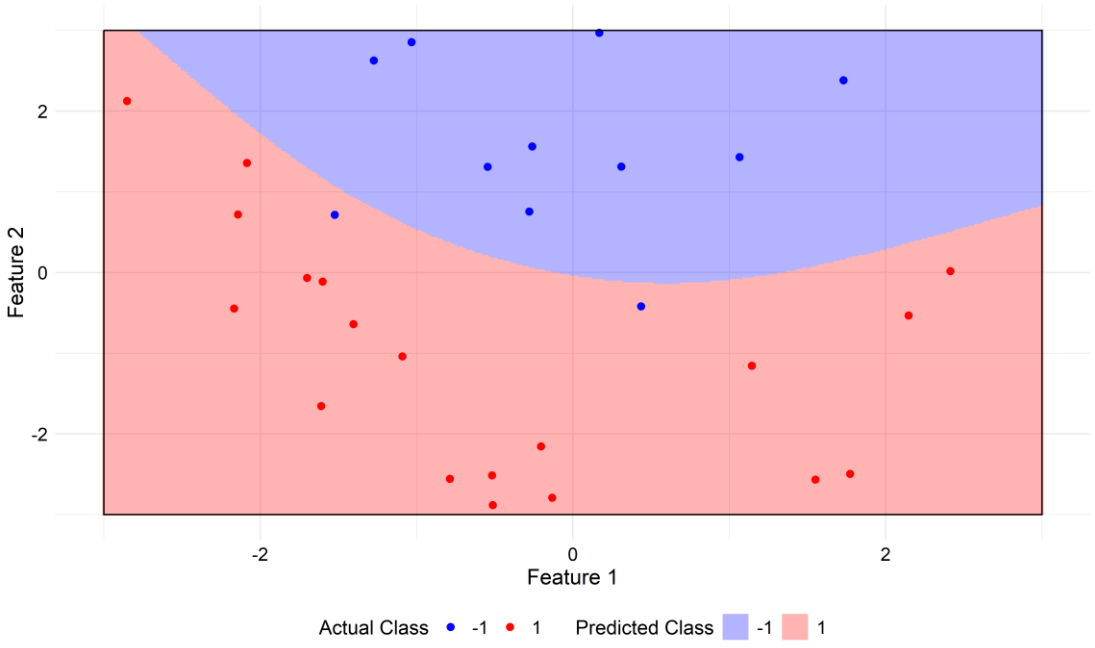

Figure 2. Nonlinear separating surface generated by a Neural Network.

2.3. eXplainable Artificial Intelligence

The so-called eXplainable Artificial Intelligence (XAI) has emerged as a critical area of research aimed at enhancing the transparency, interpretability, and trustworthiness of machine learning models (Wachter et al., 2017). XAI encompasses a diverse set of methodologies and techniques designed to elucidate the decision-making process of machine learning models, thereby facilitating human understanding and interpretability of model behavior and increasing our trust in the attained results. Among these methodologies, we highlight counterfactual explanations, feature significance analysis and sensitivity analysis.

2.3.1. Counterfactual Explanations

Counterfactual methods represent a prominent approach within the realm of XAI, focusing on the generation of alternative scenarios or 'counterfactuals' to explain model predictions. The fundamental concept underlying counterfactual methods is the creation of hypothetical instances that are similar to the observed data but differ in one or more attributes. By systematically altering the features of a given instance and observing the corresponding changes in model predictions, counterfactual methods provide valuable insights into the factors driving model decisions and predictions. This analysis typically takes the form of 'what-if' scenarios, where adjustments are made to features to generate counterfactual instances that lead to desired outcomes. By identifying the minimal changes required to alter a model prediction, counterfactual explanations shed light on the underlying decision-making process and enable decision-makers to understand the model's behavior in specific contexts. For example, in our production context the question will be 'What is the minimum amount of adjustment in inputs and/or outputs that a technically inefficient DMU would need to undertake to transition into being considered efficient at a certain probability threshold?'

2.3.2. Feature Significance Analysis and Sensitivity Analysis

To complement counterfactual analysis, it is possible to incorporate feature significance analysis, focusing on understanding the contribution of variables to the model's predictions. In this regard, several approaches exist for feature significance analysis. One approach is rule extraction methods, which aim to derive interpretable decision rules from complex models (Tickle et al., 1998; Fogel & Robinson, 2003; Martens et al., 2007). Another method involves visualization techniques, which provide graphical representations of feature influence (Craven & Shavlik, 1992; Tzeng & Ma, 2005; Cho et al., 2008). Additionally, a widely used approach is Sensitivity Analysis (SA), which assesses the impact of variable variations on model predictions (Saltelli et al., 2008; Sobol', 1993; Hamby, 1994). Lastly, more recent techniques include SHAP (Lundberg & Lee, 2017), based on cooperative game theory, and LIME (Ribeiro et al., 2016), which builds local surrogate models to approximate feature influence.

Given the variety of existing feature significance methods, we choose SA to be used in this study for several reasons. First, extraction rules typically simplify the model's complexity to produce more understandable rules, which involves discretizing the classifier, leading to information loss, and failing to accurately represent the original model. Instead, SA is a straightforward method that works with the original fitted model by systematically perturbing predictor variables and measuring the corresponding changes in the response variable, without requiring additional model retraining. Second, visualization techniques are often designed for a specific machine learning method, limiting their general applicability. This is a disadvantage compared to SA, which is model-agnostic and can be applied to any supervised machine learning method. Third, methods like SHAP and LIME pose challenges in high-dimensional datasets. SHAP, for instance, relies on Shapley values, which require exponential complexity in the number of features. Meanwhile, LIME constructs local surrogate models to approximate feature influence, but its reliance on sampling around a given instance can lead to high variance in the explanations, especially as dimensionality increases. This makes LIME's interpretability less stable in complex settings like production processes. In contrast, SA directly analyzes the original model, mitigating these issues.

3. Integrating ML techniques for classification and Data Envelopment Analysis

In this section, we perform the integration of machine learning techniques for classification tasks with Data Envelopment Analysis to enhance the measurement of technical efficiency. By combining the strengths of both methodologies, we aim to provide robust and insightful efficiency assessments of a set of DMUs. As aforementioned, while other ML classification methods could be considered, we focus on Neural Networks.

3.1. Reinterpreting DEA as a Classification Machine Learning Technique and technical efficiency measurement as a XAI method

Before introducing the methodology, we illustrate the reinterpretation of DEA as a classification method that also resorts to counterfactual analysis when measuring technical efficiency. First, following Figure 3, the DEA technology (2) differentiates the input-output space into feasible and infeasible regions. Second, DEA technical efficiency measures can be conceptualized as a classification model within the feasible category wherein the two classes represent efficient and inefficient input-output production processes, with the efficient units positioned precisely onto the boundary of the feasible-unfeasible regions, i.e., the Pareto-efficient frontier. As a classification method within the feasible region, technical efficiency measurement implies that the typical efficiency measures utilized in DEA can be reinterpreted within the realm of eXplainable Artificial Intelligence (XAI) principles, particularly in relation to the notion of counterfactual scenarios. Specifically, the movement of an inefficient DMU, by improving its observed inputs and/or outputs in accordance with the orientation and type of

efficiency measure selected (e.g., using the Directional Distance Function defined by (3)), signifies transition within its original region 'feasible' toward the exact threshold where any further minimal change would result in the unit being 'unfeasible' (through its projection onto the efficient frontier, i.e., the separating surface). This movement resembles a counterfactual that quantifies the level of technical inefficiency within the 'feasible' class through DEA (classifying firms as efficient or inefficient), thus highlighting the conceptual linkage between DEA and XAI principles.

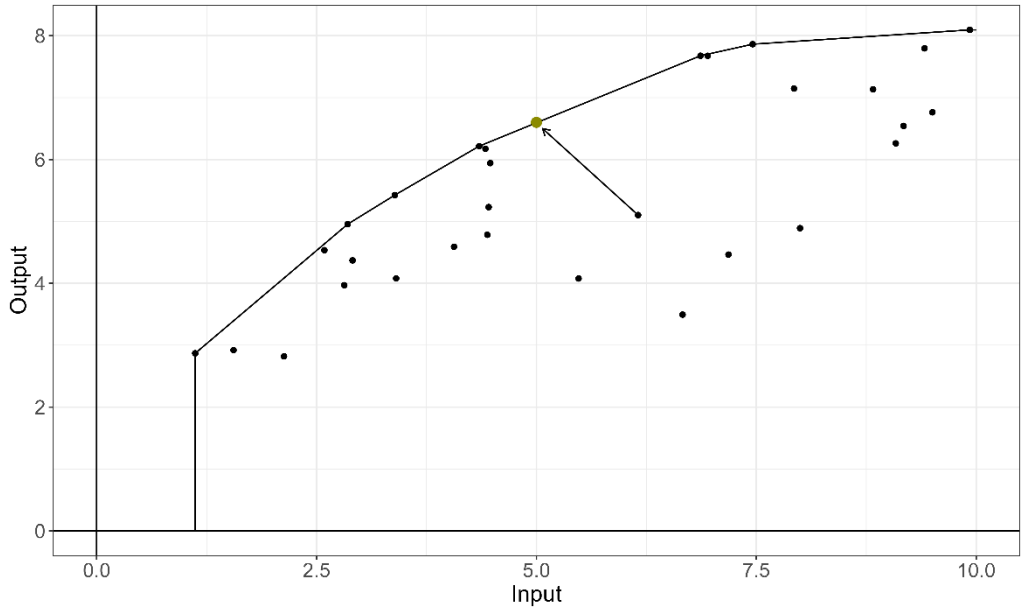

Figure 3. Data Envelopment Analysis and the Directional Distance Function.

After drawing a parallel between standard DEA approaches and classification ML methods, showing that DEA efficiency measures can be considered as a specific case of XAI, we now proceed to introduce the new methodology.

3.2. Probability-based efficiency analysis

The core concept underlying our model is a multi-stage methodology aimed at enhancing efficiency assessment through the fusion of DEA and ML techniques. The approach operates in four distinct stages: Firstly, we employ standard DEA to categorize DMUs into efficient and inefficient categories. Subsequently, in the second stage, we address the challenge of class imbalance (efficient vs. inefficient). In the third stage, we employ NNs as classification ML model, wherein the response variable is the probability of being efficient, and the classification features include both inputs and outputs. Finally, in the fourth stage we ascertain a robust measure of technical inefficiency through the application of XAI. Specifically, given a model measuring technical efficiency (specifically the DDF model (3)), we determine the required input reductions and output expansions of each inefficient DMU to transition its class from inefficient to efficient.

Next, we introduce our approach in the form of an algorithm with four steps, each corresponding to the previous stages:

Step 1 [Data labeling process]: Based on the concept of Pareto-dominance, we resort to the additive DEA model (6) below (Charnes et al., 1985) to partition the dataset of DMUs D into two categories: efficient vs. inefficient. A value of zero in the optimal solution of the linear program indicates that the evaluated DMU $\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right)$ is not Pareto-dominated by any technically feasible input-output combination within the DEA technology (2).[2] This condition underscores that for the DMU under evaluation there is no room for enhancing its input and/or output combination without compromising its technological feasibility (i.e., both input slack reductions and output slack increments are infeasible: $s_{io}^{-*} = s_{ro}^{+*} = 0$, $\forall i, r$).

$$
\begin{aligned}
A_{DEA}\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right) = \quad & \max \quad \sum_{i=1}^{m} s_{io}^{-} + \sum_{r=1}^{s} s_{ro}^{+} && && (6.0)\\
& s.t. \quad \sum_{j=1}^{n} \lambda_{jo} x_{ij} = x_{io} - s_{io}^{-}, && i = 1,\ldots,m && (6.1)\\
& \qquad \sum_{j=1}^{n} \lambda_{jo} y_{rj} = y_{ro} + s_{ro}^{+}, && r = 1,\ldots,s && (6.2)\\
& \qquad \sum_{j=1}^{n} \lambda_{jo} = 1, && && (6.3)\\
& \qquad \lambda_{jo} \geq 0, && j = 1,\ldots,n && (6.4)\\
& \qquad s_{io}^{-}, s_{ro}^{+} \geq 0, && \forall i, \forall r && (6.5)
\end{aligned}
\tag{6}
$$

If $A_{DEA}\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right) > 0$, then DMU $\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right)$ is (technically) inefficient. Otherwise, if $A_{DEA}\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right) = 0$, then DMU $\left(\boldsymbol{x}_o, \boldsymbol{y}_o\right)$ is (technically) efficient. The respective sets of efficient and inefficient DMUs are denoted as E and I. Consequently, $D = E \cup I$.

Step 2 [Class balancing phase]: Addressing the challenge of class imbalance between DMUs (efficient and inefficient) is crucial for prediction by means of ML techniques (see, for example, He & Garcia, 2009). Imbalanced datasets often compromise the performance of standard algorithms, favoring the majority class and neglecting the minority class. In our production context, large datasets typically exhibit a higher proportion of inefficient units, which can skew model outcomes and adversely affect the accuracy of predictions. To address this issue, we adopt a modified version of the Synthetic Minority

[2] The additive model can be used to measure technical efficiency (e.g., through the weighted additive model determining the inputs and output adjustments necessary to reach the frontier) instead of the Directional Distance Function (3). However, the DDF model is arguably more popular while it serves us to illustrate that efficiency measurement can be dissociated from the data labelling stage of the method. Contrarily, one could use the DDF to label the DMUs into the efficient or inefficient classes but, unfortunately, this measure does not guarantee their Pareto-efficiency categorization (as the additive model does), because individual input reductions and output increases may be feasible after the projection of the DMU to the frontier through β^* (Orea and Zofío, 2019).

Oversampling Technique (SMOTE) (Chawla et al., 2002) to generate synthetic examples of the minority class (efficient units).

However, our approach goes beyond simply balancing class proportions. Rather than just addressing class imbalance, we focus on refining the delimitation of the best-practice efficient frontier, allowing the model to learn more effectively. Depending on the dataset structure, we generate either efficient or inefficient synthetic units until the desired balance threshold is met, but not both simultaneously. If the minority class is composed of efficient units, we distribute the efficient synthetic units along the entire frontier, ensuring that the model has enough references to properly define the efficiency boundary. If additional data are still needed to reach the target proportion, we generate extra units randomly along the frontier. Conversely, if the minority class consists of inefficient units, we first calculate their inefficiency scores and then distribute the synthetic inefficient units evenly across the corresponding quartiles. This approach prevents an artificial increase in density within a specific region of the feature space—technology—while ensuring a more accurate representation of the efficient frontier. Finally, by evaluating the model's performance using a validation dataset that was not used during training, we further mitigate potential biases, as a biased model will exhibit lower performance when confronted with new data. This adaptation allows us to tailor the synthetic data generation process to better fit the characteristics of our dataset and context. Next, we describe the specific implementation process of our adapted approach to generate synthetic units.

First, we determine the necessary number of synthetic units to balance the proportion of units in both classes (efficient vs. inefficient). Since there is no universally optimal ratio, Weiss and Provost (2003) suggest testing different minority proportions to identify the most effective distribution for the training set. Considering the minimization of the classification error, they conclude that the ideal proportion of minority class should fall between 20% and 40%. Following this approach. We treat this proportion as a hyperparameter of the ML model, denoted by $\pi_{\min} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$. As aforementioned, we adopt two different strategies to achieve this set of balances depending on the dataset's observed class distribution and the selected minority class proportion $\pi_{\min}$.

- *Case 2a (minority class: efficient DMUs):* As the proportion of the originally observed efficient DMUs in the observed dataset (*D*) does not reach the balance level $\pi_{\min}$, it is necessary to increase their number by creating additional synthetic efficient DMUs, whose set is denoted by $\hat{E}$. Specifically, we generate convex combinations from sets of $m+s$ DMUs labeled as efficient in Step 1. The objective of this approach is to populate complete faces of the convex polyhedron. The total number of combinations is calculated as $\binom{n_E}{m+s}$, where n_E is the number of efficient DMUs labeled as efficient in Step 1 and $m+s$ is the total number of variables considered in the problem.

For each combination of observed efficient DMUs in the $m+s$ dimensions, a synthetic unit is generated by applying the same weights to all DMUs involved in that linear combination. The weight is defined as $\upsilon = \frac{1}{m+s}$, chosen arbitrarily to position the synthetic units at the midpoint of the convex combination. By choosing efficient DMUs in all input and output dimensions, $m+s$, we ensure that the reference frontier used to create the synthetic DMUs corresponds to full-dimensional facets. Once all convex combinations have been created, we use the additive DEA model (6) to identify which of these combinations are Pareto-efficient. If the number of efficient units remains insufficient, additional random DMUs are generated. In this process, the weights for the observed efficient DMUs are randomly selected within the open interval (0,1). To maintain consistency and ensure that the sum of all weights equals 1, each weight is normalized by dividing it by the total sum of all weights, yielding a new relative weight for each DMU. When the pre-fixed balance level is achieved, the generation of synthetic units stops. If none of the $\binom{n_E}{m+s}$ combinations yield a Pareto-efficient point, we proceed by selecting combinations of $m+s-1$ efficient DMUs. If this approach also fails, we reduce the number to $m+s-2$ and continue iterating in this manner until a solution is found. The whole process results in the balanced dataset $\hat{D} = D \cup \hat{E} = E \cup I \cup \hat{E}$ with the desired proportion $\pi_{\min}$.

– *Case 2b (minority class: inefficient DMUs):* Now it is necessary to create additional synthetic inefficient DMUs ($\hat{I}$). The process consists of four stages. First, as in the previous case, synthetic convex combinations from efficient DMUs are obtained using equal weights. However, rather than identifying and keeping those combinations that are efficient, we select those that are inefficient. Therefore, in the second stage the additive DEA model (6) is needed to determine which of these combinations are inefficient. Third, and separately, a large random sample of convex combinations of the initially identified inefficient DMUs is generated—for example, 20 times the desired number of synthetic inefficient units. Fourth, based on a quantile choice (e.g., quartiles, quintiles, etc.) of the originally observed distribution of inefficiency scores, $A_{DEA}(\boldsymbol{x}_o, \boldsymbol{y}_o) > 0$, equally numbered subsets of synthetic units by quantiles are randomly selected until the targeted balance proportion is achieved. This systematic approach ensures a well-distributed (populated) set of synthetic inefficient DMUs by maintaining representativeness across different inefficiency levels. The resulting balanced dataset at the selected proportion $\pi_{\min}$ is now $\hat{D} = D \cup \hat{I} = E \cup I \cup \hat{I}$.

Step 3 [Fitting the ML model]: In this phase, a classification-based ML model—NN in this paper—is implemented, where the dependent variable denotes the efficiency status (efficient [class +1] vs. inefficient [class −1]), while the predictive variables (features) comprise all inputs and outputs. Additionally, model parameters are fine-tuned at this stage. Specifically, these parameters are

determined by the minority class proportion $(\pi_{\min})$ and the hyperparameters of the selected NN methodology. These parameters correspond to the model fitting size (γ), representing the number of neurons in the hidden layer, and the decay parameter (α), corresponding to the intensity with which the neuronal weights are updated in each iteration of the back-propagation algorithm.

The training phase consists of two stages. In the first stage, we apply the NN model using *k*-fold cross-validation on the extended datasets $\hat{D}$ (for each $\pi_{\min}$ level), to determine the optimal parameter configuration (γ,α) for the model, i.e., $\pi_{\min}\left(\gamma^*,\alpha^*\right)$. In the second stage, we compare the performance of these best-trained models to identify the ideal balance proportion between efficient and inefficient DMUs. This results in the best classification model $\Gamma\left(\boldsymbol{x},\boldsymbol{y};\pi^*_{\min}\left(\gamma^*,\alpha^*\right)\right)$.

To measure the performance of each specific balance proportion in the second stage, we rely on standard metrics commonly used in ML classification problems. We consider measures that not only pay attention to the two classes like 'Balanced accuracy' (the average of true predictions for each class), but also that are primarily focused on metrics related to the 'efficient' (minority) class of interest, such as 'Precision' (the proportion of positive predictions that are actually correct), 'Sensitivity' (i.e., the proportion of actual positives correctly identified) and, finally, 'F1-score' (the harmonic mean of 'Precision' and 'Sensitivity', balancing detection accuracy and reliability). Depending on the specific context and the metrics of interest, we evaluate model performance by ranking the results according to our chosen criteria.

If the original dataset is sufficiently large, we propose creating training and validation partitions to evaluate the model's performance with data not used during the fitting phase. This approach helps mitigate overfitting and provides a reliable estimate of the model's real-world performance when applied to new data. However, it is quite common that datasets are too small to allow for such partitioning (as in our empirical application). In this event we suggest testing the models trained with different balance levels on the originally observed (i.e., unbalanced) dataset, D. This ensures that all models are evaluated on the same dataset, allowing for a fair performance comparison. If multiple models achieve the same performance according to the above metrics on the original dataset, we propose to reevaluate only the tied models based on their performance using their respective balanced datasets, $\hat{D}$, whose sizes vary depending on the balance level. Finally, if equality persists, the smallest $\pi_{\min}$ level is selected, following the principle of parsimony (as fewer synthetic observations are required to reach $\pi_{\min}$). A toy example in the following section will illustrate the process.

Finally, once the best classification model $\Gamma\left(\boldsymbol{x},\boldsymbol{y};\pi^*_{\min}\left(\gamma^*,\alpha^*\right)\right)$ is determined, it is used to predict the probability of being technically efficient. Given the predicted probability *p* (according with eq. (5))

of any input-output bundle $(\boldsymbol{x}, \boldsymbol{y})$—including observed DMUs, it is classified as technically efficient (+1) or inefficient (–1) following the standard NN rule: if $\Gamma\left(\boldsymbol{x}, \boldsymbol{y}; \pi^*_{\min}\left(\gamma^*, \alpha^*\right)\right) > 0.5$, then $(\boldsymbol{x}, \boldsymbol{y})$ is classified as efficient; otherwise, $(\boldsymbol{x}, \boldsymbol{y})$ is classified as inefficient.

Step 4 [Measuring technical inefficiency]: The final step is to measure technical efficiency using a standard technical efficiency measure. For instance, the DDF presented in model (3), whose the directional vector corresponds to $\left(\boldsymbol{g}^x, \boldsymbol{g}^y\right)$. In our approach, each component of this vector is interpreted as the relative weight assigned to inputs and outputs when measuring inefficiency. In particular, we define each element of $\boldsymbol{g}^x$ as the weight capturing the relative importance of input *i* multiplied by its respective observed mean, and each element of $\boldsymbol{g}^y$ as the relative importance of output *r* multiplied by its respective observed mean. The relative importance of each variable is calculated through the SA method (see Section 2.3.2). As we are aware, the selection of this type of directional vector is also original (Wang et al., 2019). This choice of directional vector is particularly noteworthy for several reasons. First, incorporating the relative importance of input and output variables—derived from their contribution to predicting efficiency probability—introduces a novel and objective criterion for defining the direction of improvement. Unlike traditional approaches that often rely on subjective or uniform weightings, this method exploits data-driven insights obtained through sensitivity analysis. Second, using the same directional vector to evaluate inefficiency across all DMUs ensures that the inefficiency values—optimal β^*'s—remain comparable across units. This uniformity enhances the interpretability of inefficiency scores, as differences in inefficiency are not influenced by DMU-specific directional choices but rather reflect genuine performance gaps relative to a common yardstick. Third, by constructing the directional vector based on the mean values of inputs and outputs in the original data set *D*, we ensure that the measure is inherently dependent on the original units of measurement. This property makes the inefficiency measure units' invariant, meaning that results remain consistent regardless of the scale or units in which the inputs and outputs are expressed. This avoids potential distortions that could arise from differences in variable magnitudes or measurement scales.

Next, the value of technical inefficiency β^* is determined using counterfactual analysis. In particular, we answer the question: "What is the minimum modification required for the evaluated DMU to be classified as efficient with at least probability $\bar{p}$?". The corresponding projection (input-output targets) is calculated through (7) where $\boldsymbol{x}_o$ and $\boldsymbol{y}_o$ are the observed inputs and outputs for DMU_o, respectively, and β takes values within a predefined grid:

$$\left(\boldsymbol{x}_o - \beta \boldsymbol{g}^x, \boldsymbol{y}_o + \beta \boldsymbol{g}^y\right). \tag{7}$$

For each β in the grid, we determine $\Gamma^*\left(\boldsymbol{x}_o-\beta\boldsymbol{g}^x,\boldsymbol{y}_o+\beta\boldsymbol{g}^y;\ \pi^*_{\min}\left(\gamma^*,\alpha^*\right)\right)$. If the target probability $\bar{p}$ lies between two consecutive β values, we refine the search by iteratively testing new β values within that range and recalculating their associated probabilities $\Gamma^*\left(\boldsymbol{x}_o-\beta\boldsymbol{g}^x,\boldsymbol{y}_o+\beta\boldsymbol{g}^y;\ \pi^*_{\min}\left(\gamma^*,\alpha^*\right)\right)$. When the algorithm converges to the probability that meets the desired confidence threshold, the corresponding β^* value is recorded as the minimum adjustment required for the DMU to achieve technical efficiency.

Moreover, given a predefined probability threshold denoted by $\bar{p}$, the benchmark for each evaluated DMU is identified as the closest unit—based on the Euclidean distance—classified as efficient at level $\bar{p}$. This selection ensures that the chosen peer represents the most comparable efficient unit, providing a meaningful benchmark for performance evaluation. By resorting to proximity within the input-output space, the approach facilitates a more intuitive interpretation of efficiency adjustments required for the assessed DMU to reach the predefined efficiency level. Note that peers can change for each DMU depending on $\bar{p}$. Formally, for DMU $\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)$, and given probability threshold $\bar{p}$, its corresponding peer is determined as $\left(\boldsymbol{x}_{j^*},\boldsymbol{y}_{j^*}\right)=\arg\min_{j=1,\ldots,n}\left\{\left\|\left(\boldsymbol{x}_o,\boldsymbol{y}_o\right)-\left(\boldsymbol{x}_j,\boldsymbol{y}_j\right)\right\|_2:\Gamma\left(\boldsymbol{x}_j,\boldsymbol{y}_j;\ \pi^*_{\min}\left(\gamma^*,\alpha^*\right)\right)\geq\bar{p}\right\}$. In case of a tie, we propose reporting the group of peers for evaluation and choosing one at random.

Additionally, if in the evaluation process, the probability of being efficient of a DMU exceeds the pre-fixed threshold $\bar{p}$, its β^* is set to 0 as it is 'superefficient' at the selected probability level. This implies that we do not calculate negative β^* values related to increasing inputs and reducing outputs to project DMUs 'backwards' towards lower-level probability frontiers—although it would be possible following Andersen and Petersen (1993). Finally, it is worth noting that when searching for the potential benchmarks, our algorithm does not consider projections with input values smaller than the minimum observed across the sample, as it could result in negative valued inputs, which is unrealistic. As a result, some DMUs could never reach the established threshold $\bar{p}$. In such cases, we consider β^* the biggest feasible β in the considered grid.

3.3. An illustrative example

Next, we illustrate the novel methodology through a numerical example. We consider a simulated dataset comprising 40 DMUs, $D=\left\{\left(x_j,y_j\right)\right\}_{j=1}^{40}$, where each DMU_j employs a single input to produce a single output. In accordance with the proposed algorithm, Step 1 assigns efficiency labels by solving the additive DEA model (6). Specifically, DMU_j is classified as efficient if and only if all its optimal

slacks are zero. In our simulated dataset, only 3 DMUs are Pareto-efficient, namely DMUs 6, 7 and 31 (see Figure 4), while the remaining DMUs are inefficient. Consequently, the observed label distribution is highly imbalanced, with approximately 7.5% of the DMUs classified as efficient and 92.5% as inefficient.

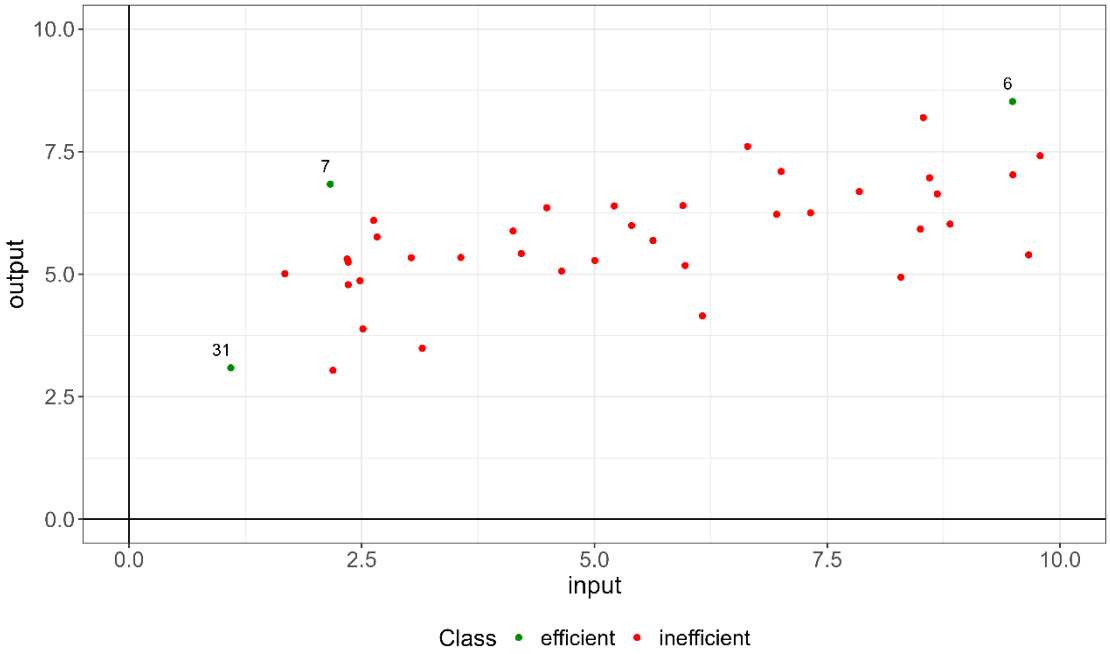

Figure 4. Labeling through the standard DEA additive model.

Step 2 addresses the dataset imbalance by creating synthetic observations for the efficient minority class in this example. Synthetic efficient units, denoted by $\hat{E}$, are generated as linear combination of the 3 efficient DMUs until the total number of efficient units reaches the target proportion $\pi_{\min}$. This augmentation results in a new dataset, $\hat{D} = D \cup \hat{E} = E \cup I \cup \hat{E}$. We create the necessary synthetic observations to achieve the proposed range of minority classes $\pi_{\min} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$. Figure 5 illustrates the case where the efficient units represent 25% of the observations. This balance percentage corresponds to the outcome obtained through the optimizing balancing process obtained in the third step, resulting in the best NN classification model $\Gamma\left(\boldsymbol{x}, \boldsymbol{y};\ \pi^*_{\min}\left(\gamma^*, \alpha^*\right)\right)$.

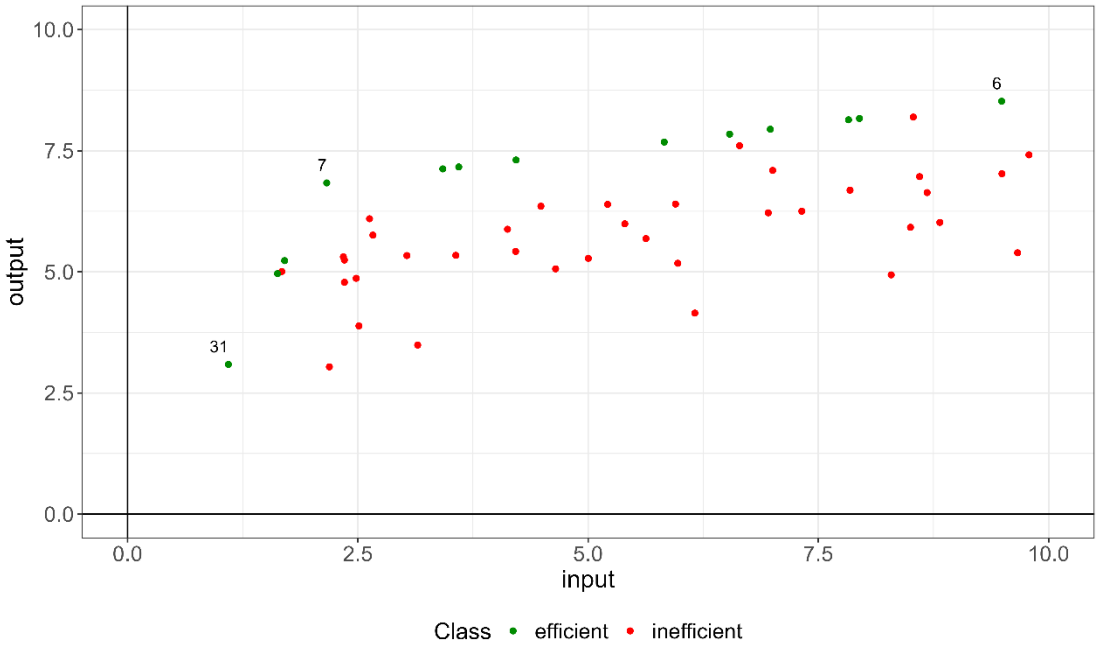

Figure 5. The labeled dataset that will be used for model training.

Step 3 involves training the machine learning model. Seeking simplicity, we employ a NN with an initial (input) layer containing as many nodes as input and output variables (represented in in Figure 1 by I1 (input) and I2 (output)), a single hidden layer (H1), and a final exit (output) layer consisting of a single neuron (O1). In practice, we use the *R* package '*caret*' (Kuhn, 2008) to facilitate model training, specifically using the NN implementation from the '*nnet*' package (Venables & Ripley, 2002).

Once the different levels of the minority class $\pi_{\min} \in \{0.2, 0.25, 0.3, 0.35, 0.4\}$ have been defined and their associated extended datasets constructed, i.e., $\hat{D}(\pi_{\min}) = D \cup \hat{E}(\pi_{\min})$, we consider for each value in $\pi_{\min}$ a grid of possible values for the two hyperparameters: (i) model fitting size: $\gamma \in \{1, 5, 10, 20, 30\}$, and (ii) decay parameter: $\alpha \in \{0,\ 0.1,\ 0.01,\ 0.001,\ 0.0001\}$. To determine the optimal hyperparameters in the first stage, we perform a 5-fold cross-validation. For each fold we randomly use 80% of the observations to train the NN and the remining 20% for testing. After that, we evaluate the performance of each hyperparameter configuration (γ, α) for each minority class proportion $\pi_{\min}$ using the observed dataset D. Once the optimal value of the hyperparameters is obtained for each balance level, $\pi_{\min}(\gamma^*, \alpha^*)$ in the second stage we evaluate the performance of the different balance levels. Due to the limited sample size of D, we do not use in this example a separate validation subset and use the whole observed dataset.

Table 1 ranks the performance of the fitted models for different $\pi_{\min}(\gamma^*, \alpha^*)$, using the metrics discussed in the previous section. To rank the models, we propose the following order. First, using

'Balanced accuracy', we select the model with the highest accuracy rate in predicting the two classes. If multiple models achieve the same score, we then evaluate them based on metrics that prioritize the minority 'efficient' class. Specifically, we use the 'F1-score', as it balances detection and confidence through the harmonic mean. If a tie persists, we further assess models based on confidence using 'Precision', and finally, on detection using 'Sensitivity'. In this example, performance for all metrics is tied for minority class distributions of 25%, 35%, and 40% (see upper panel of Table 1). Consequently, since all previous metrics are equal, we further reevaluate performance using the augmented dataset $\left(\hat{D}\right)$ for the tied balance levels, which include both real and synthetic units—i.e., the training dataset. After this evaluation, configurations with minority class proportion of 25% and 40% remain tied (see lower panel of Table 1. Performance using train dataset). Hence, following the principle of parsimony, we select the dataset with a minority class proportion of 0.25.

Table 1. Performance results for each fitted model across different class balancing levels, ranked from highest to lowest performance.

Performance using observed dataset (*D*)

$\pi_{\min}$	Balance accuracy	F1-score	Precision	Sensitivity
0.25	1	1	1	1
0.35	1	1	1	1
0.4	1	1	1	1
0.3	0.99	0.86	0.75	1
0.2	0.83	0.8	1	0.67

Performance using training dataset ($\hat{D}$) (only tied balance levels)

$\pi_{\min}$	Balanced accuracy	F1	Precision	Sensitivity
0.25	1	1	1	1
0.4	1	1	1	1
0.35	0.99	0.98	0.95	1

The above procedure identifies the best in-class NN probability model, corresponding to $\Gamma\left(\boldsymbol{x},\boldsymbol{y};\ 0.25,20,0\right)$. Figure 6a (top) presents the results of the classification model. The observed DMUs are represented as black points. Additionally, two distinct regions are identified choosing $\bar{p}$ = 0.5 as the probability threshold to differentiate among the efficient and inefficient classes. When visualizing the classification results, the probability threshold used to separate the two classes is flexible and can be modified by the user, directly impacting the displayed areas for each class. In this case we identify the green region with input/output production processes to which the model assigns a probability greater than 0.5, classifying these units as "efficient", and the pink region, where the probability is 0.5 or lower, classifying units as "inefficient". Moreover, Figure 6b (bottom) represents uncertainty through a gradient based on the probabilities predicted by the model. Areas representing

units with a predicted probability between 0.25 and 0.75 are shaded in black, indicating high uncertainty about their efficient or inefficient status, while regions with probabilities closer to 0 or 1 are displayed in white, reflecting greater certainty about their class.

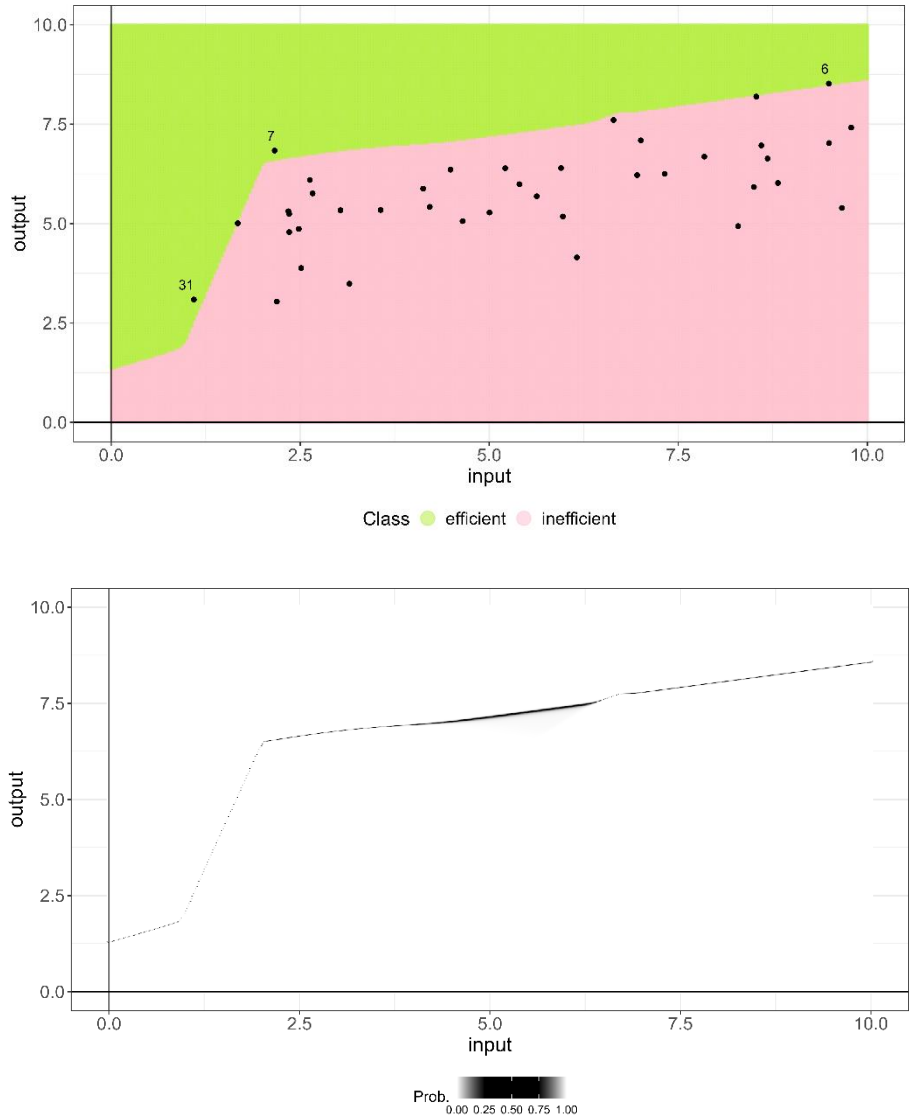

Figure 6a (*Top*). *Predicted efficient/inefficient regions generated by the new approach are displayed alongside the original unlabeled DMUs. Figure 6b* (*Bottom*). *Uncertainty regions are shaded in black as predicted by the fitted model, with certainty regions shown in white.*

Once the final model is established, we use it to measure technical efficiency in step 4. First, we perform a SA analysis (see Section 2.3.2) using the '*Rminer*' library (Cortez, 2010) to determine the directional projection vector for the DDF. Specifically, the SA results are $SA_x = 0.333$ for the input and $SA_y = 0.667$ for the output. This analysis reveals that the model assigns twice as much importance to the output variable as to the input variable when classifying a DMU as efficient or inefficient. We then

define the directional vector as $(\boldsymbol{g}_x, \boldsymbol{g}_y) = (SA_x \cdot \bar{x}, SA_y \cdot \bar{y}) = (1.804, 3.848)$, where $\bar{x}$ and $\bar{y}$ represent the mean value of the input x and output y in the observed data set *D*.

Lastly, we determine β^* (the efficiency score derived from the DDF), along with the input and output targets and the peers for each DMU. In particular, we illustrate the measurement of inefficiency at a 0.75 confidence level. Figure 7 displays the separating surface, highlighting DMUs 6, 7 and 31—the three DEA efficient DMUs— with a probability of being efficient greater than 0.75, and located in the green area ($\bar{p}$ >0.75). Additionally, the DDF projection of DMU 22, classified as inefficient with an input value of 4.49 and an output value of 6.36, is shown. Using the previously defined directional vector, we calculate the value of β^* required for DMU 22 to reach the specified efficiency confidence level ($\bar{p}$ =0.75). As a result, its projection reduces the input to 4.17 (its input target) and increases the output to 7.03 (its output target), corresponding to a $\beta_{22}^* = 0.17$. From a benchmarking perspective, DMU 22 has DMU 7 as its peer at a 0.75 confidence level. This selection is based on the Euclidean distance, where DMU 7 is the closest among the originally observed efficient units at a 0.75 probability level (6, 7, and 31) to the projection of DMU 22 (point 22').

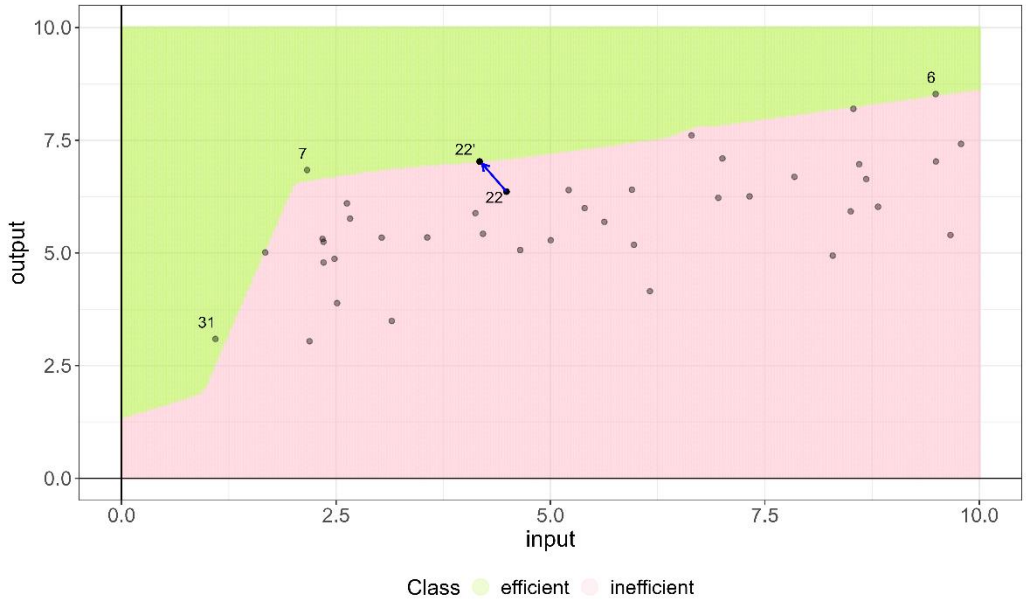

Figure 7. Projection of DMU 22 at an efficiency confidence level of $\bar{p}$ *=0.75.*

4. An empirical application: Efficiency Assessment of the Valencian Food Industry

In this section, we show that the new approach can be rendered operational using real-world data of firms operating in the Spanish food industry, a sector that plays a crucial role in the country's economy. The food industry in Spain is both economically significant and culturally rich, seamlessly combining traditional practices with modern technological innovations. Its scope covers the entire food value

chain, transforming raw agricultural products into a wide variety of food items consumed domestically and exported internationally. This industry is supported by a diverse ecosystem, ranging from small-scale farmers dedicated to preserving heritage techniques to large companies adopting advanced industrial production systems. It is also a vital source of employment, spanning agriculture, processing, distribution, and retail. Numerous studies worldwide have analyzed efficiency in food industries, with examples including India (Kumar and Basu, 2008), Mexico (Flegl et al., 2022), Taiwan (Dadura and Lee, 2011), and Indonesia (Machmud et al., 2019). Such analyses provide valuable insights into the operational dynamics of food sectors across different regions.

In Spain, the institutional and economic environment is shaped by its 17 autonomous communities, each characterized by distinct policies and market conditions. This regional diversity introduces significant complexity into any analysis, as variations in regulations and management frameworks influence business operations at the community level. The Valencian Community (located in the eastern coast of Spain by the Mediterranean Sea), selected as the focus of this study, exemplifies such diversity. Known for its strong agricultural exports and medium-sized enterprises, this region provides a representative case for evaluating efficiency in the Spanish food industry.

The dataset used for this analysis consists of 97 food industry companies located in the Valencian Community, each employing more than 50 workers, and collected from the SABI database for year 2023.[3] The dataset includes several variables that comprehensively reflect the operational and financial profiles of the companies. The output variable, operating income (in millions of Euros), captures revenue generated from core business activities. Input variables include total assets (in millions of Euros), representing the resources utilized; the number of employees, indicating workforce size; tangible fixed assets (in millions of Euros), such as buildings and machinery essential for production; and personnel expenses (in millions of Euros), encompassing costs like salaries, benefits, and training. Together, these variables enable a detailed examination of resource allocation, labor engagement, and financial investments, forming the basis for a robust analysis of operational efficiency within the Valencian food industry (see Table 2). To better understand the characteristics of the dataset and the challenges it presents for analysis, Table 2 presents the descriptive statistics for the sample. Examining the data, we observe that the dataset includes both very small and very large companies. The maximum and minimum values are significantly distant from the mean and median, highlighting the wide dispersion in the data. This dispersion affects the central tendency measures, resulting in a notable difference between the mean and median.

Building upon this production framework, we employ the technique described in this paper, which combines ML techniques for classification and DEA, to determine a robust technical efficiency analysis.

[3] The SABI (Iberian Balance Sheet Analysis System) database is a subset of the ORBIS product commercialized by Moody's, which offers comparable data on private companies: https://sabi.bvdinfo.com.

This approach allows us to capture the complex intricacies and idiosyncrasies of the food industry in the Valencian Community, providing a more accurate and contextualized perspective on efficiency.

Table. 2 Summary statistics of input and output variables.

	Inputs				Output
	Total assets	Employees	Fixed assets	Personnel expenses	Operating income
Min.	1.537	50	0.142	1.037	2.382
1st Qu.	8.989	75	2.680	2.167	12.994
Median	24.555	98	6.258	3.059	29.138
Mean	41.030	201	15.280	6.757	62.307
3rd Qu.	52.409	240	20.096	8.244	72.688
Max.	258.825	1076	140.689	36.789	460.578

Source: SABI (Iberian Balance Sheet Analysis System) database.

4.1. Balancing the classes in the dataset

In this dataset $D=\left\{\left(x_i, y_i\right)\right\}_{i=1}^{97}$, the additive model (6) identifies 15 out of the 97 firms as efficient. After classifying the data, the next step involves balancing the two classes to create the training dataset, $\hat{D}=D\cup\hat{E}$ in this case, and fine-tune the NN. Regarding the balance of classes, since 15.43% of the firms are labeled as efficient, which is a value close to the minimum recommended percentage of 20%, we also compare the performance of this model to scenarios where the efficient class is balanced between 20% and 40% at increments of 5%. In each scenario, we apply a NN with a unique hidden layer. For each proposed proportion for the minority class $\pi_{\min}\in\{0.20,\ 0.25,\ 0.30,\ 0.35,\ 0.40\}$, a grid of selected hyperparameters is defined for model fitting, including model fitting size $\gamma\in\{1,\ 5,\ 10,\ 20,\ 30\}$, and decay parameter $\alpha\in\{0,\ 0.1,\ 0.01,\ 0.001,\ 0.0001,\ 0.00001\}$. We follow the same process for selecting the best-performing NN model as described in Section 3.2 and the example. After selecting the best parameter configurations at each balance level $\pi_{\min}\left(\gamma^*,\alpha^*\right)$, we find the optimal classification NN, evaluating their performance with 'Balanced accuracy' first, followed by the 'F1-score', 'Precision', and 'Sensitivity'. Table 3 presents the performance of each parameter configuration, with the optimal results obtained by $\pi_{\min}^*=0.4$, $\gamma^*=20$ and $\alpha^*=0.01$, i.e., the best probability NN model is $\Gamma\left(\boldsymbol{x},\boldsymbol{y};\ \pi_{\min}^*\left(\gamma^*,\alpha^*\right)\right)$. Notably, the second-best performance in terms of 'Balanced accuracy' is observed when the dataset is not balanced, with a minority class distribution of 15.43%. However, this model has the worst 'Precision' (0.60) while achieving the highest possible 'Sensitivity' (1), indicating that it tends to overclassify firms as efficient, resulting in less reliable predictions.

For this dataset, we see that balancing the classes improves precision at the expense of sensitivity when comparing the model trained on the original dataset *D* to other models trained on balanced datasets $\hat{D}$, as observed in the datasets with 20% and 40% balance, where 'Precision' reaches 1. Regarding 'Sensitivity', the unbalanced model tends to overclassify units as efficient, as we previously mentioned. However, as we introduce balancing, this issue diminishes, achieving a very good performance across all metrics at 20% balance. At this balance level, 'Balanced Accuracy' remains nearly the same, but the maximization of precision makes the model's predictions more reliable. Nevertheless, further increasing the minority class proportion seems to introduce confusion into the model, making classification more challenging. At 25%, 30%, and 35% balance levels, the performance declines compared to the 20% balance dataset. However, when the dataset reaches a minority class percentage of 40%, the detection of true positives (sensitivity) increases again to 0.93. Additionally, 'Precision' reaches 1, meaning that all observations detected by the model are correctly classified in the validation set, with no false positives. The NN trained on the dataset with 40% balance achieves the best performance across all metrics, making it the most consistent model regardless of the evaluation criteria, including Balanced Accuracy, F1-Score, Precision, and Sensitivity.

Table. 3 Performance results of different models depending on balancing levels, ranked by performance.

Performance using observed dataset *D*.

$\pi_{\min}$	Balanced accuracy	F1-Score	Precision	Sensitivity
0.40	0.97	0.97	1	0.93
0.1543*	0.94	0.75	0.60	1
0.20	0.93	0.93	1	0.87
0.30	0.85	0.79	0.85	0.73
0.35	0.85	0.79	0.85	0.73
0.25	0.79	0.69	0.82	0.60

Note: * Observed balance proportion in *D*.

4.2. Efficiency probabilities and sensitivity analysis

After selecting the best-performing model, we predict through the fitted NN the probability of being efficient for each firm. Table 4 ranks the firms in the sample based on their efficiency probabilities. We report on the top 25 firms, with the last firm showing already a very low probability value, which indicates that the remaining 69 firms are classified as inefficient with certainty. As shown in the third column (probability of being efficient), a total of 14 DMUs are predicted to have a probability of efficiency exceeding 0.5, one less than initially labelled in the first step by standard DEA (model (6)).

Table 4. Top 25 DMUs ranked by the probability of being efficient thresholds and their corresponding peers at different thresholds.

Ranking	Firm	Probability of being efficient	$\bar{p}$ = 0.75			$\bar{p}$ = 0.85			$\bar{p}$ = 0.95		
			β^*	Reached Probability	Peer	β^*	Reached Probability	Peer	β^*	Reached Probability	Peer
1	2	0.9999	0	0.75	2	0	0.85	2	0	0.95	2
2	18	0.9998	0	0.75	18	0	0.85	18	0	0.95	18
3	3	0.9996	0	0.75	3	0	0.85	3	0	0.95	3
4	17	0.9983	0	0.75	17	0	0.85	17	0	0.95	17
5	20	0.9962	0	0.75	20	0	0.85	20	0	0.95	20
6	36	0.9960	0	0.75	36	0	0.85	36	0	0.95	36
7	46	0.9894	0	0.75	46	0	0.85	46	0	0.95	46
8	1	0.9868	0	0.75	1	0	0.85	1	0	0.95	1
9	56	0.9705	0	0.75	56	0	0.85	56	0	0.95	56
10	62	0.9486	0	0.75	62	0	0.85	62	0	0.9486	46
11	93	0.9441	0	0.75	93	0	0.85	93	0	0.9441	46
12	92	0.9335	0	0.75	92	0	0.85	92	0	0.9335	46
13	9	0.9288	0	0.75	9	0	0.85	9	0.8734	0.95	3
14	97	0.9176	0	0.75	97	0	0.85	97	0	0.9176	46
15	25	0.4981	0.3033	0.75	17	0.3650	0.85	17	0.4815	0.95	17
16	26	0.4909	0.4131	0.75	17	0.5025	0.8078	17	0.5025	0.8078	17
17	91	0.0735	0.1005	0.5385	93	0.1005	0.5385	93	0.1005	0.5385	56
18	22	0.0560	0.2665	0.7500	18	0.2942	0.85	18	0.3480	0.95	18
19	43	0.0549	0.1005	0.4121	46	0.1005	0.4121	46	0.1005	0.4121	46
20	85	0.0490	0.0994	0.75	93	0.1005	0.7596	93	0.1005	0.7596	46
21	95	0.0338	0.1986	0.7500	56	0.2145	0.85	56	0.2495	0.95	56
22	83	0.0335	0	0.0335	92	0	0.0335	92	0	0.0335	46
23	44	0.0183	0.0710	0.75	56	0.0824	0.85	56	0.1091	0.95	56
24	75	0.0167	0	0.0167	62	0	0.0167	62	0	0.0167	46
25	94	0.0099	0	0.0099	93	0	0.0099	93	0	0.0099	46

Next, we perform SA on the selected model. The relative importance of the variables used by the model to assign efficiency probabilities is presented in Figure 8. 'Operating income' accounts for 50.8% of the total importance, highlighting its dominant role in the model's decision-making process. Among the input variables, the relative importance is distributed as follows: 'Employees' (25.4%), 'Total assets' (12.5%), 'Fixed assets' (11.2%), and 'Personnel expenses' (0.1%). These relative importance results are subsequently used to define the directional vector as $(5.129, 51.054, 1.711, 0.007, 31.652)$ and calculate the β^* value (the DDF) for each DMU through counterfactual analysis, as explained in Section 3.

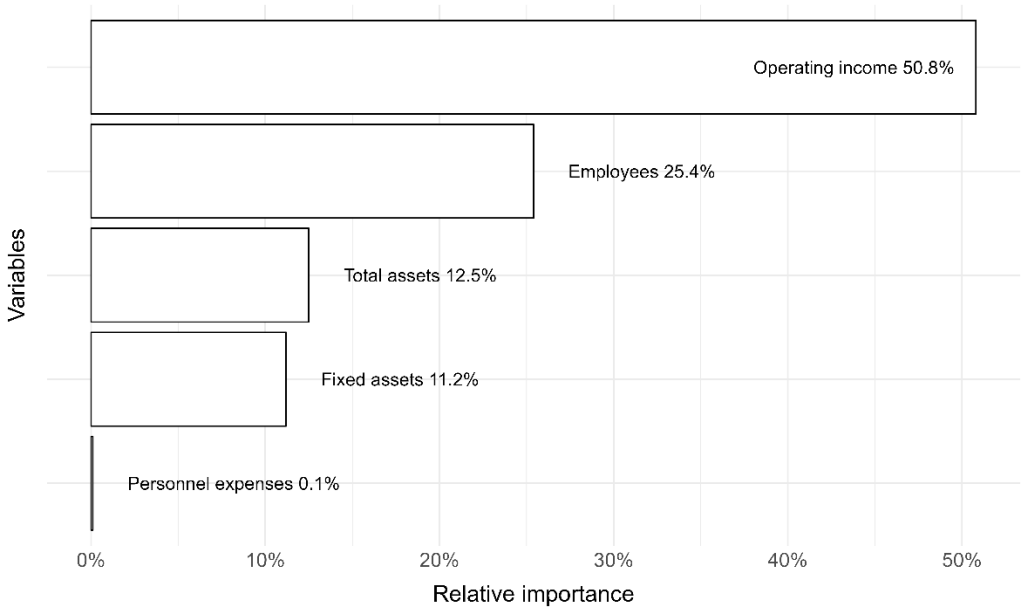

Figure 8. Relative importance of variables, ordered by significance.

4.3. Technical efficiency and benchmark peers

We report in Table 4 the inefficiency values and benchmark peers for each firm using three thresholds corresponding to $\bar{p} = 0.75$, $\bar{p} = 0.85$ and $\bar{p} = 0.95$. For example, the top 9 firms are efficient at the 95% level and therefore their β^*'s are equal to zero. Decreasing the asked probability threshold results in more firms being classified as efficient. The first 14 firms exhibit probabilities very close to 1, being efficient at $\bar{p} = 0.90$. Meanwhile, firms ranked 15th and 16th can be labeled inefficient because their efficiency probabilities do not reach the threshold $\bar{p} = 0.5$ by less than 1 percentage point. The remaining firms show probabilities close to 0. Notably, firm #26 (ranked 16th and discussed below), originally labelled as efficient by the additive DEA model (6), is now classified as inefficient, demonstrating the ability of NNs to overcome the deterministic nature of the traditional approach by using the probabilistic paradigm associated to ML classifying models. In the peer columns, we indicate the reference benchmark for each firm at each probability threshold. If a firm is efficient at a given threshold, i.e., $\beta^* = 0$, it is its own peer. Considering the highest scenario $\bar{p} = 0.95$, all inefficient firms identify their reference peer in the top 9, all of which have probabilities exceeding 0.95. Across the entire dataset, all peers belong to the 14 firms that are efficient with a probability greater than 0.5.

We now discuss in depth the inefficiency profile of four firms (#22, #26, #83 and #36) and their projections for a scenario with a confidence level of $\bar{p}$ =0.85. These firms are selected as representative examples of different efficiency profiles and to illustrate the varying output and input adjustments required to reach the probabilistic frontier.

- **Firm #22**, ranked 18th in Table 4, represents a standard case where the evaluated firm presents increasing inefficiency scores as the demanded probability threshold increases: $\beta^*(\bar{p}=0.75) = 0.27$, $\beta^*(\bar{p}=0.85) = 0.29$, and $\beta^*(\bar{p}=0.95) = 0.35$, implying that to reach the probabilistic frontiers, inputs need to be reduced and the output increased to an ever increasing extent. In Table 5, we present the observed inputs and output as well as the predicted ones corresponding to the projections on the frontier at the selected probability level $\bar{p} = 0.85$. We also report the difference between the two as the percentage changes in the inputs and output required to reach the probability frontier. Finally, the closest reference peer to such projection is identified; in this case firm #18 (the 2nd firm in the ranking), which remains the same at the three probability thresholds.
- **Firm #26**, ranked 16th, portraits an interesting case. First, it was classified as efficient by the standard DEA additive model (6), while the model reports a probability of being efficient $p = 0.49$. Second, while it can improve the probability of being efficient by decreasing inputs and increasing outputs, it fails to meet the threshold $\bar{p} = 0.85$. The maximum probability that it can reach with an efficiency score $\beta^* = 0.5$, is lower than the pre-defined threshold: i.e., $\bar{p} > p^{\max}(\beta^* = 0.5) = 0.81$ (see Tables 4 and 5). At this threshold level, the unit cannot decrease its inputs anymore because it reaches the minimum observed value of 50 employees. That is, for higher betas, the predicted value of employees would be lower than the minimum observed in *D* (added as additional restriction to model (3)), thereby preventing unrealistic predictions that would signal resource levels close to zero (or even negative). The reference peer is firm #17 both at $\bar{p} = 0.75$ and $p^{\max}(\beta^* = 0.5) = 0.81$.
- **Firm #83**, ranked 22nd, exemplifies the case of a firm that cannot increase its efficiency because, as before, it uses almost the minimum observed amount of 50 employees, and therefore there is no room for improvement, $\beta^* = 0$. Despite not being able to increase its efficiency probability, i.e., $p^{\max}(\beta^* = 0) = 0.03$, we identify that the closest peer at the $p = 0.85$ probability frontier is firm #92.
- **Firm #36**, ranked 6th, represents the opposite case. This firm is efficient with probability $p = 0.99$, well above the three reference thresholds considered in Tables 4 and 5, thereby having $\beta^* = 0$ and being its own peer.

Overall, the adjustments (resource reallocations) required for more efficient firms (smaller betas or higher probabilities of being efficient) are smaller than for less efficient ones (greater betas or smaller probabilities of being efficient). For instance, the adjustments for firm #22 ($\beta^* = 0.29$) are notably smaller than those of firm #26 ($\beta^* = 0.50$), both in absolute values and percentage terms (in brackets). However, despite being closer to the probability frontier $\bar{p} = 0.85$, the efficiency probability of firm #22 is smaller because the model is certain about its inefficient status. On the contrary, firm #26 is

further away from the probability frontier, but the model is uncertain about its status assigning an efficiency probability of $p = 0.49$. Therefore, two firms could be at the same distance of the probability frontier (i.e., have equal betas), but the efficiency probability could be different depending on the certainty/uncertainty assigned by the NN model to its efficiency status. As shown in Figure 6b, this depends on the relative location of the input-output bundle with respect to the probability frontier trained on the balanced dataset (i.e., the separating surface between the efficient and inefficient regions which can be blurry). This underlines the importance of populating the imbalanced class to obtain more reliable predictions. An ideal situation would unambiguously classify observations as efficient and inefficient with probability values close to one and zero, respectively, which would reflect the lack of uncertainty. Then, the efficiency scores would truly reflect the changes in the input and output quantities necessary to reach the efficient class. However, this would make it impossible to study the changes necessary to reach different probability thresholds $\bar{p}$, which are the result of the existence of some level of uncertainty.

Table 5. Efficiency profiles of firms 22, 26, 83, and 36 for $\bar{p}$ =0.85.

	Firm #22			Firm #26		
	Peer (#18)			Peer (#17)		
	Observed	Predicted		Observed	Predicted	
Total assets	24.71	23.20	(-6.11%)	27.93	25.36	(-9.23%)
Employees	212	196.97	(-7.09%)	80	54.33	(-32.09%)
Fixed assets	11.46	10.96	(-4.39%)	13	12.14	(-6.62%)
Personnel expenses	8	8.00	(-0.02%)	2.17	2.16	(-0.16%)
Operating income	80.89	90.21	(11.51%)	72.57	88.47	(21.92%)
Probability, p	0.056	$\bar{p}$ =0.85		0.49	p^{max} = 0.81	
β^*	-	0.29		-	0.50	
	Firm 83			Firm 36		
	Peer (#92)			Peer (#36)		
	Observed	Projection		Observed	Predicted	
Total assets	7.24	7.24	(0%)	67.59	67.59	(0%)
Employees	51	51	(0%)	53	53	(0%)
Fixed assets	0.66	0.66	(0%)	5.21	5.21	(0%)
Personnel expenses	2.04	2.04	(0%)	3.91	3.91	(0%)
Operating income	8.89	8.89	(0%)	54.91	54.91	(0%)
Probability, p	0.03	p^{max} = 0.03		0.99	$\bar{p}$ =0.85	
β^*	-	0		-	0	

This discussion shows that even if the measurement of technical efficiency is contingent on the efficiency probabilities predicted by the NN, the efficiency scores cannot be inversely related to the probability of being efficient (i.e., smaller betas would correspond to higher efficiency probabilities and

vice versa). The absence of a monotonic relationship suggests that to rank observations one could consider first if the firm can reach the desired probability threshold $\bar{p}$, followed by the maximum projected efficiency probability that can be attained $p^{\max}(\beta^*)$ and, finally, the probability of being efficient. Annex 1 reports this ranking, complementary to that presented in Table 4, and considering $\bar{p}$ = 0.85 as probability threshold. The first 14 firms exceed this threshold with individual probabilities $p > \bar{p} = 0.85$, being classified as efficient and ranked according with the probability of being efficient, Subsequently, firms ranked 15th to 66th with $p < p^{\max}(\beta^*) = \bar{p} = 0.85$, are capable of reaching the probability threshold by reducing their inputs and increasing their output in the amount given by their efficiency scores β^*, which are used to rank the firms in increasing order, i.e., the smaller the beta, the greater the rank, and showing that highly ranked firms need to adjust less their production processes. Finally, from firm #26 (ranked 67th) onwards, firms are ranked by the maximum probability that they can achieve given their efficiency scores, i.e., $p \leq p^{\max}(\beta^*) < \bar{p} = 0.85$. For instance, firms #40 and #8 (ranked 68th and 69th) have a common maximum probability of $p^{\max}(\beta^*) = 0.80$. But the former has a smaller beta, so it is ranked first.

4.4. Efficient resource reallocation at the industry level

Besides analyzing the efficiency profiles of individual firms, we are interested in computing the main statistics for the efficient projections in each scenario to analyze the overall input and output adjustments required to reach the frontier. For the thresholds $\bar{p} \in \{0.75,\ 0.85,\ 0.95\}$, a total of 25, 31, and 37 firms, respectively, fail to achieve the established probability level, $p^{\max}(\beta^*) < \bar{p}$, due to the constraints discussed in Section 3—specifically, the restriction that prevents reducing any input below the minimum observed value in each dimension. To further analyze these results, Tables 6, 7, and 8 summarize the mean, median, and standard deviation for these scenarios, highlighting the impact of increasing probability thresholds on the projections (in particular, on the input and output targets). The percentage increments are shown in brackets for better interpretation. According to Table 6, if all firms were to adjust the processes following the proposed directional vector to be considered efficient, total assets would, on average, decrease by 13% (for $\bar{p}$ =0.75) to 15% (for $\bar{p}$ =0.95), employees by 26% to 30%, and fixed assets by 12% to 13%. Operating income would increase by 53% to 61%, depending on the scenario, while personnel expenses remain unchanged, reflecting their negligible role in the adjustments resulting from the sensitivity analysis. On average, the probability of being efficient in the observed dataset is 0.15, rising to 0.6, 0.67, and 0.74 in the predicted scenarios. These results depict realistic situations where some firms cannot reach the efficient frontier for high thresholds, leading the sector to converge toward a specific probability level below 1. Additionally, average β values progressively increase to 1.03, 1.14, and 1.20, reflecting the growing effort required to achieve higher

confidence thresholds. Note that since we are using a common directional vector for all observations, the efficiency scores can be consistently aggregated (employing equal unitary weights) and averaged (through the arithmetic mean), being representative of the industry performance, Briec et al. (2009).

While Table 6 highlights average adjustments, Table 7 focuses on median values, which offer a more representative view of the input and output changes required for the typical firm to reach the efficiency thresholds. For the median firm, total assets would decrease by 22% to 23%, employees by 26% to 31%, and fixed assets by 4% to 5%. Operating income would increase by 52% to 61%, consistent with the mean results. As before, personnel expenses remain unaffected. The larger reductions in total assets observed in the median case suggest that smaller or more typical firms require relatively greater adjustments to achieve efficiency compared to larger DMUs that influence the sector average. Additionally, Table 7 reveals a significant increase in the median probability of efficiency, progressing from 0 in the observed dataset to 0.75, 0.85, and 0.95, underscoring the adaptability of the model. Finally, Table 8 reflects the variability in the input reductions and output increases necessary to reach the different efficiency thresholds with respect to the observed values. Compare to the observed data, standard deviations decrease at the projections, but remain similar despite the probability threshold.

Table 6. Mean values of observed data and projections at different confidence levels, with percentage changes from observed values shown in parentheses.

Scenario	Observed	$\bar{p}$ =0.75		$\bar{p}$ =0.85		$\bar{p}$ =0.95	
Total assets	41.03	35.72	(-13%)	35.18	(-14%)	34.86	(-15%)
Employees	201.00	148.29	(-26%)	142.90	(-29%)	139.70	(-30%)
Fixed assets	15.28	13.51	(-12%)	13.33	(-13%)	13.22	(-13%)
Personnel expenses	6.76	6.75	(0%)	6.75	(0%)	6.75	(0%)
Operating income	62.31	95.05	(53%)	98.39	(58%)	100.37	(61%)
Probability *p*	0.15	0.60		0.67		0.74	
Beta β^*	0.00	1.03		1.14		1.20	

Table 7. Median values of observed data and projections at different confidence levels, with percentage changes from observed values shown in parentheses.

Scenario	Observed	$\bar{p}$ =0.75		$\bar{p}$ =0.85		$\bar{p}$ =0.95	
Total assets	24.56	19.12	(-22%)	19.05	(-22%)	18.88	(-23%)
Employees	98.00	72.66	(-26%)	69.87	(-29%)	68.06	(-31%)
Fixed assets	6.26	6.03	(-4%)	6.01	(-4%)	5.97	(-5%)
Personnel expenses	3.06	3.05	(0%)	3.05	(0%)	3.05	(0%)
Operating income	29.14	44.35	(52%)	45.47	(56%)	46.90	(61%)
Probability *p*	0.00	0.75		0.85		0.95	
Beta β^*	0.00	0.38		0.40		0.49	

Table 8. Standard deviation of observed data and projections at different confidence levels, with percentage changes from observed values shown in parentheses.

Scenario	Observed	$\bar{p}$ =0.75		$\bar{p}$ =0.85		$\bar{p}$ =0.95	
Total assets	49.75	45.01	(-10%)	44.25	(-11%)	44.14	(-11%)
Employees	216.03	162.32	(-25%)	157.61	(-27%)	156.09	(-28%)
Fixed assets	23.26	21.03	(-10%)	20.63	(-11%)	20.61	(-11%)
Personnel expenses	7.69	7.68	(0%)	7.68	(0%)	7.68	(0%)
Operating income	84.17	115.49	(37%)	121.50	(44%)	122.34	(45%)
Probability, p	0.34	0.28		0.32		0.36	
Beta β^*		1.65		1.90		1.91	

Overall, we conclude form Tables 6 and 7 that that as the probability threshold (confidence level) of belonging to the efficient class increases, the required adjustments in inputs and outputs for the evaluated firms to be classified as efficient also increase. Therefore, as expected, a higher probability threshold demands a greater magnitude of change, meaning that the evaluated firms must undergo more significant input and output reallocations to achieve efficiency. This result highlights the increasing effort required to meet stricter efficiency classification criteria.

5. Conclusions and future work

A growing body of literature explores the integration of Machine Learning (ML) with Data Envelopment Analysis (DEA) to enhance efficiency analysis across various sectors. While many studies have focused on improving traditional DEA methodologies through ML techniques based on regression, our research extends this synergy by introducing classification models to predict efficiency probabilities. Our findings show that the integration of ML classifiers with DEA can predict the efficiency status of Decision-Making Units (DMUs) while offering a richer framework for assessing efficiency through probabilistic measures and counterfactual analysis. The advantages of our integrated approach extend beyond just analytical improvements. They also offer practical benefits in terms of scalability and adaptability. The model's ability to handle large datasets efficiently makes it especially relevant in the era of big data, where organizations across sectors are looking to exploit vast amounts of information for enhanced decision-making (Zhu, 2022). Additionally, the flexibility of the ML-DEA framework means it can be tailored to specific sector needs. This novel approach is illustrated through an empirical analysis of food industry firm data, emphasizing its practical utility.

As a summary, let us highlight that the new approach introduces several key methodological, interpretative, and practical contributions to efficiency analysis by integrating ML techniques within a DEA framework. First, we propose a novel classification-based from Tables 6 and 7 approach in the second stage of a ML-DEA hybrid framework, moving beyond traditional regression-based techniques. In the first stage, we employ a standard DEA model to generate a binary efficiency label, which is then

predicted in the second stage using classification models. Second, our framework enhances inferential power by estimating the probability of a DMU being classified as efficient, shifting DEA from a purely descriptive tool to a probabilistic efficiency assessment. This aligns efficiency analysis with modern inferential analytics and decision-making frameworks. Third, the novel approach allows reframing DEA as a classification problem, where the efficiency frontier is understood as a separating surface between technically feasible and infeasible input-output profiles, allowing efficiency measures to be framed in terms of the minimal modifications required for reclassification. Fourth, our approach is algorithm-agnostic, enabling robust efficiency assessments across various classification models, including decision trees, SVMs, neural networks, and ensemble methods. Fifth, we integrate Explainable AI (XAI) techniques, particularly counterfactual analysis, to define inefficiency in terms of the minimum changes required for an inefficient DMU to become efficient, offering an interpretable efficiency assessment. Sixth, we introduce a benchmarking approach that exploits the importance ranking of inputs and outputs obtained with machine learning models of sensitivity analysis that allow assigning data-driven weights to directional projections, thereby improving the interpretability and strategic value of efficiency assessments. Seventh, we enhance benchmarking by incorporating probabilistic efficiency thresholds, allowing for target setting through counterfactual benchmarking, and offering actionable improvement strategies based on minimum necessary input-output modifications. Eighth, we propose a new ranking system for DMUs based on their probabilistic efficiency scores, offering an alternative to traditional DEA ranking methods. Finally, our method facilitates a refined proximity-based benchmark identification strategy, ensuring that each DMU is compared against the closest efficient benchmark at any given efficiency probability threshold, strengthening the practical applicability of DEA for dynamic and adaptive benchmarking. These contributions collectively advance efficiency analysis by bridging the gap between DEA, statistical learning, and explainable AI, offering a more flexible and interpretable approach to performance assessment.

Looking forward, several research avenues appear promising. First. the exploration of other machine learning techniques, such as ensemble methods (e.g., Random Forest or Boosting) could provide further improvements in the robustness and accuracy of efficiency predictions. Indeed, when faced with a real empirical case, we could implement multiple machine learning techniques in parallel (not only NN) to assess the consistency and robustness of the results. By comparing the outcomes across different models, we could evaluate the stability of efficiency classifications and ensure that our findings are not overly dependent on a specific algorithm. Secondly, the application of our integrated ML-DEA model to other domains, such as environmental sustainability, public sector performance, or benchmarking through composite indicators could be highly beneficial. These areas, where efficiency and resource optimization are critical, may significantly benefit from the enhanced analytical capabilities that our model offers. Additionally, extending our model to handle real-time data could transform operational efficiency monitoring, allowing organizations to make immediate adjustments

based on current performance metrics. Lastly, further research should also focus on the development of more sophisticated counterfactual methods within the ML-DEA framework.

References

Amirteimoori. A. Allahviranloo. T. Zadmirzaei. M. & Hasanzadeh. F. (2023). On the environmental performance analysis: a combined fuzzy data envelopment analysis and artificial intelligence algorithms. *Expert Systems with Applications*. *224*. 119953.

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. Management science, 39(10), 1261-1264.

Aparicio, J., Esteve, M., Rodriguez-Sala, J. J., & Zofio, J. L. (2021). The estimation of productive efficiency through machine learning techniques: efficiency analysis trees. In *Data-enabled analytics: DEA for big data* (pp. 51-92). Cham: Springer International Publishing.

Aydin. N. & Yurdakul. G. (2020). Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms. *Applied Soft Computing*. *97*. 106792.

Banker. R. D. & Morey. R. C. (1986). Efficiency analysis for exogenously fixed inputs and outputs. Operations Research. 34(4). 513-521.

Banker. R. D. Charnes. A. & Cooper. W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*. *30*(9). 1078-1092.

Berger. A. N. Brockett. P. L. Cooper. W. W. & Pastor. J. T. (1997). New approaches for analyzing and evaluating the performance of financial institutions. *European Journal of Operational Research*. *98*(2). 170-174.

Boubaker. S. Le. T. D. Ngo. T. & Manita. R. (2023). Predicting the performance of MSMEs: A hybrid DEA-machine learning approach. *Annals of Operations Research*. 1-23.

Briec, W Dervaux B. & Leleu H. (2009). Aggregation of directional distance functions and industrial efficiency. Journal of Economics, 79, 237–61.

Chambers, R. G., Chung, Y., & Färe, R. (1998). Profit, directional distance functions, and Nerlovian efficiency. Journal of optimization theory and applications, 98, 351-364.

Charles. V. Aparicio. J. & Zhu. J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. European Journal of Operational Research. 279(3). 929-940.

Charnes. A. Cooper. W. W. & Rhodes. E. (1978). Measuring the efficiency of decision making units. European Journal of Operational Research. 2(6). 429-444.

Charnes. A. Cooper. W. W. Golany. B. Seiford. L. & Stutz. J. (1985). Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of econometrics*. *30*(1-2). 91-107.

Chawla. N. V. Bowyer. K. W. Hall. L. O. & Kegelmeyer. W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 16. 321-357.

Chen. Y. Li. Y. Xie. Q. An. Q. & Liang. L. (2014). Data envelopment analysis with missing data: a multiple imputation approach. International Journal of Information and Decision Sciences. 6(4). 315-337.

Cho. B. H. Yu. H. Lee. J. Chee. Y. J. Kim. I. Y. & Kim. S. I. (2008). Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine*. *12*(2). 247-256.

Cortez. P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. In *Industrial conference on data mining* (pp. 572-583). Berlin. Heidelberg: Springer Berlin Heidelberg.

Craven. M. W. & Shavlik. J. W. (1992). Visualizing learning and computation in artificial neural networks. *International journal on artificial intelligence tools*. *1*(03). 399-425.

Dadura, A. M., & Lee, T. R. (2011). Measuring the innovation ability of Taiwan's food industry using DEA. *Innovation: The European Journal of Social Science Research*, *24*(1-2), 151-172.

Daouia. A. Noh. H. & Park. B. U. (2016). Data envelope fitting with constrained polynomial splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. *78*(1). 3-30.

Emrouznejad. A. & Shale. E. (2009). A combined neural network and DEA for measuring efficiency of large scale datasets. *Computers & Industrial Engineering*. *56*(1). 249-254.

Esteve. M. Aparicio. J. Rabasa. A. & Rodriguez-Sala. J. J. (2020). Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees. Expert Systems with Applications. 162. 113783.

Esteve. M. Aparicio. J. Rodriguez-Sala. J. J. & Zhu. J. (2023). Random Forests and the measurement of super-efficiency in the context of Free Disposal Hull. European Journal of Operational Research. 304(2). 729-744.

Fallahpour. A. Olugu. E. U. Musa. S. N. Khezrimotlagh. D. & Wong. K. Y. (2016). An integrated model for green supplier selection under fuzzy environment: application of data envelopment analysis and genetic programming approach. *Neural Computing and Applications*. *27*. 707-725.

Flegl, M., Jiménez-Bandala, C. A., Sánchez-Juárez, I., & Matus, E. (2022). Analysis of production and investment efficiency in the Mexican food industry: Application of two-stage DEA. *Czech Journal of Food Sciences*, *40*(2).

Fogel. D. B. & Robinson. C. J. (2003). Techniques for extracting classification and regression rules from artificial neural networks.

Goodfellow. I. Bengio. Y. & Courville. A. (2016). Deep Learning. MIT Press.

Guerrero. N. M. Aparicio. J. & Valero-Carreras. D. (2022). Combining Data Envelopment Analysis and Machine Learning. Mathematics 2022. 10. 909.

Guillen. M. D. Aparicio. J. & España. V. J. (2023a). boostingDEA: A boosting approach to Data Envelopment Analysis in R. *SoftwareX*. *24*. 101549.

Guillen. M. D. Aparicio. J. & Esteve. M. (2023b). Gradient tree boosting and the estimation of production frontiers. *Expert Systems with Applications*. *214*. 119134.

Guillen. M. D. Aparicio. J. & Esteve. M. (2023c). Performance Evaluation of Decision-Making Units Through Boosting Methods in the Context of Free Disposal Hull: Some Exact and Heuristic Algorithms. *International Journal of Information Technology & Decision Making*. 1-30.

Guillen. M. D. Aparicio. J. Zofío. J. L. & España. V. J. (2024). Improving the predictive accuracy of production frontier models for efficiency measurement using machine learning: The LSB-MAFS method. Computers & Operations Research. 171. 106793.

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. Environmental Monitoring and Assessment, 32(2), 135-154.

He. H. & Garcia. E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 21(9). 1263-1284.

Jin. Q. Kerstens. K. & Van de Woestyne. I. (2024). Convex and nonconvex nonparametric frontier-based classification methods for anomaly detection. OR Spectrum. 1-27.

Jomthanachai. S. Wong. W. P. & Lim. C. P. (2021). An application of data envelopment analysis and machine learning approach to risk management. *Ieee Access*. *9*. 85978-85994.

Khademian, A. (2024). Optimization of blasting patterns in Esfordi phosphate mine using hybrid analysis of data envelopment analysis and multi-criteria decision making. *Engineering Applications of Artificial Intelligence*, *133*, 108061.

Kuhn. Max (2008). "Building Predictive Models in R Using the caret Package." *Journal of Statistical Software*. **28**(5). 1–26.

Kumar, M., & Basu, P. (2008). Perspectives of productivity growth in Indian food industry: a data envelopment analysis. *International Journal of Productivity and Performance Management*, *57*(7), 503-522.

Kuosmanen. T. & Johnson. A. L. (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research*. *58*(1). 149-160.

Kwon. H. B. Lee. J. & Roh. J. J. (2016). Best performance modeling using complementary DEA-ANN approach: Application to Japanese electronics manufacturing firms. *Benchmarking: An International Journal*. *23*(3). 704-721.

LeCun. Y. Bengio. Y. & Hinton. G. (2015). Deep learning. Nature. 521(7553). 436-444.

Liao. Z. Dai. S. & Kuosmanen. T. (2024). Convex support vector regression. *European Journal of Operational Research*. *313*(3). 858-870.

Lin. S. W. & Lu. W. M. (2024). Using inverse DEA and machine learning algorithms to evaluate and predict suppliers' performance in the apple supply chain. *International Journal of Production Economics*. 109203.

Liu. H. H. Chen. T. Y. Chiu. Y. H. & Kuo. F. H. (2013). A comparison of three-stage DEA and artificial neural network on the operational efficiency of semi-conductor firms in Taiwan. *Modern Economy*. *4*(01). 20-31.

Lundberg. S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Machmud, A., Ahman, E., Dirgantari, P. D., Waspada, I., & Nandiyanto, A. B. D. (2019). Data envelopment analysis: The efficiency study of food industry in Indonesia. *Journal of Engineering Science and Technology*, *14*(1), 479-488.

Martens. D. Baesens. B. Van Gestel. T. & Vanthienen. J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*. *183*(3). 1466-1476.

Nandy. A. & Singh. P. K. (2020). Farm efficiency estimation using a hybrid approach of machine-learning and data envelopment analysis: Evidence from rural eastern India. *Journal of Cleaner Production*. *267*. 122106.

Olesen. O. B. & Ruggiero. J. (2022). The hinging hyperplanes: An alternative nonparametric representation of a production function. *European Journal of Operational Research*. *296*(1). 254-266.

Olesen. O. B. Petersen. N. C. & Podinovski. V. V. (2007). Staff assessment and productivity measurement in public administration: an application of data envelopment analysis. Omega. 35(3). 297-307.

Omrani. H. Emrouznejad. A. Teplova. T. & Amini. M. (2024). Efficiency evaluation of electricity distribution companies: Integrating data envelopment analysis and machine learning for a holistic analysis. Engineering Applications of Artificial Intelligence. 133. 108636.

Orea, L. & Zofío, J.L. (2019). Common methodological choices in non-parametric and parametric analyses of firms' performance, in ten Raa, T. & Greene, W. H. (eds.) *Palgrave Handbook of Economic Performance Analysis*, Palgrave-MacMillan: Cham, 419-484.

Parmeter. C. F. & Racine. J. S. (2013). Smooth constrained frontier analysis. *Recent Advances and Future Directions in Causality. Prediction. and Specification Analysis: Essays in Honor of Halbert L. White Jr*. 463-488.

Pastor. J. T. Lovell. C. K. & Aparicio. J. (2012). Families of linear efficiency programs based on Debreu's loss function. *Journal of Productivity Analysis*. *38*. 109-120.

Ribeiro. M. T. Singh. S. & Guestrin. C. (2016. August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Sahil, M. A., Tiwari, A., & Lohani, Q. D. (2025). Two-stage type-2 fuzzy parabolic double frontier data envelopment analysis. *Engineering Applications of Artificial Intelligence*, *144*, 110154.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global Sensitivity Analysis: The Primer. Wiley.

Sexton, T. R. (1986). The methodology of data envelopment analysis. New directions for program evaluation, 32, 7-29.

Soboĺ, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.*, *1*, 407.

Tayal. A. Solanki. A. & Singh. S. P. (2020). Integrated framework for identifying sustainable manufacturing layouts based on big data. machine learning. meta-heuristic and data envelopment analysis. *Sustainable Cities and Society*. *62*. 102383.

Thanassoulis, E., Portela, M. C., & Despic, O. (2008). Data envelopment analysis: the mathematical programming approach to efficiency analysis. The measurement of productive efficiency and productivity growth, 251-420.

Thanassoulis. E. Boussofiane. A. & Dyson. R. G. (2015). Applied data envelopment analysis. Springer.

Tickle. A. B. Andrews. R. Golea. M. & Diederich. J. (1998). The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Transactions on Neural Networks. 9(6). 1057-1068.

Tsionas. M. Parmeter. C. F. & Zelenyuk. V. (2023). Bayesian artificial neural networks for frontier efficiency analysis. *Journal of Econometrics*. *236*(2). 105491.

Tzeng. F. Y. & Ma. K. L. (2005). *Opening the black box-data driven visualization of neural networks* (pp. 383-390). IEEE.

Valero-Carreras. D. Aparicio. J. & Guerrero. N. M. (2021). Support vector frontiers: A new approach for estimating production functions through support vector machines. *Omega*. *104*. 102490.

Valero-Carreras. D. Aparicio. J. & Guerrero. N. M. (2022). Multi-output support vector frontiers. *Computers & Operations Research*. *143*. 105765.

Valero-Carreras. D. Moragues. R. Aparicio. J. & Guerrero. N. M. (2024). Evaluating different methods for ranking inputs in the context of the performance assessment of decision making units: A machine learning approach. Computers & Operations Research. 163. 106485.

Venables W.N. & Ripley B.D. (2002). *Modern Applied Statistics with S*. Fourth edition. Springer. New York. ISBN 0-387-95457-0.

Wachter. S. Mittelstadt. B. & Russell. C. (2017). 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.' Harvard Journal of Law & Technology. 31(2). 841-887.

Wang, K., Xian, Y., Lee, C. Y., Wei, Y. M., & Huang, Z. (2019). On selecting directions for directional distance functions in a non-parametric framework: a review. *Annals of Operations Research*, *278*, 43-76.

Weiss. G. M. & Provost. F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. Journal of artificial intelligence research. 19. 315-354.

Zhou. P. Ang. B. W. & Poh. K. L. (2008). A survey of data envelopment analysis in energy and environmental studies. European Journal of Operational Research. 189(1). 1-18.

Zhu. J. (2022). DEA under big data: Data enabled analytics and network data envelopment analysis. Annals of Operations Research. 309(2). 7

Zhu. N. Zhu. C. & Emrouznejad. A. (2021). A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies. *Journal of Management Science and Engineering*. *6*(4). 435-448.

Annex. Ranking of firms sorted by target probability and efficiency score β^*.

Ranking efficiency	Firm	Target: $\bar{p}$ = 0.85 Reached Probability	β^*	Peer	Probability of being efficient
1	2	0.85	0	2	0.9999
2	18	0.85	0	18	0.9998
3	3	0.85	0	3	0.9996
4	17	0.85	0	17	0.9983
5	20	0.85	0	20	0.9962
6	36	0.85	0	36	0.996
7	46	0.85	0	46	0.9894
8	1	0.85	0	1	0.9868
9	56	0.85	0	56	0.9705
10	62	0.85	0	62	0.9486
11	93	0.85	0	93	0.9441
12	92	0.85	0	92	0.9335
13	9	0.85	0	9	0.9288
14	97	0.85	0	97	0.9176
15	44	0.85	0.0824	56	0.0183
16	15	0.85	0.1454	18	0.0023
17	88	0.85	0.197	93	0.007
18	95	0.85	0.2145	56	0.0338
19	42	0.85	0.2633	36	0.0038
20	71	0.85	0.2765	56	0.0005
21	64	0.85	0.2873	56	0.0018
22	31	0.85	0.2896	18	0.0001
23	22	0.85	0.2942	18	0.056
24	73	0.85	0.2981	93	0.0001
25	69	0.85	0.3097	56	0.0017
26	67	0.85	0.3202	56	0.0021
27	76	0.85	0.3259	93	0.0029
28	25	0.85	0.365	17	0.4981
29	68	0.85	0.3714	56	0.0004
30	58	0.85	0.3951	46	0.005
31	86	0.85	0.4024	56	0.0019
32	48	0.85	0.4458	46	0.0001
33	53	0.85	0.4476	56	0
34	61	0.85	0.4492	56	0.0001
35	51	0.85	0.5399	56	0.0006
36	49	0.85	0.5584	18	0.0009
37	55	0.85	0.6641	56	0.0005
38	41	0.85	0.8121	18	0.0006
39	60	0.85	0.8146	56	0.0006
40	84	0.85	1.0532	56	0.0012
41	12	0.85	1.092	18	0.0013
42	79	0.85	1.1175	56	0.0006

43	57	0.85	1.1426	18	0.0009
44	30	0.85	1.1637	18	0.0005
45	47	0.85	1.2221	56	0.0006
46	45	0.85	1.2521	18	0.0006
47	4	0.85	1.329	9	0.0014
48	19	0.85	1.4364	9	0.0008
49	35	0.85	1.4846	18	0
50	13	0.85	1.4867	9	0.0008
51	39	0.85	1.8448	18	0.0006
52	24	0.85	1.847	18	0.0001
53	10	0.85	2.0748	9	0.0008
54	29	0.85	2.1556	18	0.0006
55	38	0.85	2.4489	9	0.0006
56	34	0.85	2.5875	18	0.0006
57	21	0.85	2.6141	9	0
58	16	0.85	3.1932	9	0
59	32	0.85	3.2558	18	0.0006
60	23	0.85	3.3215	9	0.0006
61	33	0.85	4.2378	9	0.0006
62	14	0.85	4.7428	9	0.0006
63	7	0.85	5.5861	9	0.0001
64	11	0.85	5.7558	9	0
65	6	0.85	7.4255	9	0.0006
66	5	0.85	8.2035	9	0
67	26	0.8078	0.5025	17	0.4909
68	40	0.8007	2.3116	18	0.0001
69	8	0.8007	11.6583	9	0
70	37	0.8006	1.608	56	0.0006
71	27	0.7995	2.1106	18	0.0001
72	85	0.7596	0.1005	93	0.049
73	72	0.74	0.3015	93	0.0001
74	70	0.6879	0.201	62	0.0026
75	63	0.5543	0.9045	56	0.0006
76	91	0.5385	0.1005	93	0.0735
77	90	0.4433	0.3015	56	0.0085
78	54	0.4247	0.5025	46	0.0001
79	43	0.4121	0.1005	46	0.0549
80	65	0.0776	0.5025	46	0.0006
81	52	0.0741	2.0101	56	0
82	28	0.0739	0.603	20	0.0003
83	50	0.0527	0.5025	46	0
84	83	0.0335	0	92	0.0335
85	59	0.0225	0.1005	62	0.001
86	74	0.0212	0.5025	46	0
87	75	0.0167	0	62	0.0167
88	94	0.0048	0.1005	93	0.0099

89	82	0.0023	0	92	0.0023
90	81	0.0021	0	18	0.0021
91	89	0.0018	0.1005	93	0.0027
92	96	0.0012	0	92	0.0012
93	80	0.0009	0	92	0.0009
94	78	0.0006	0	92	0.0006
95	66	0.0004	0.1005	62	0.0006
96	77	0.0001	0.201	93	0.0005
97	87	0.0001	0.5025	46	0