International Journal of Operations Research and Artificial Intelligence

www.ijorai.reapress.com

Int. J. Oper. Res. Artif. Intell. Vol. 1, No. 1 (2025) 20–28.

Paper Type: Original Article

Integrating DEA with Machine Learning for Predictive Modeling in Breast Cancer

Zahra Bagheri[1,*], Faranak Taghizadeh Firoozabadi[2]

[1] Department of Mathematics, Khomeinishahr Branch, Islamic Azad University, Isfahan, Iran; zahra66bagheri@gmail.com.

[2] Department of Mathematics, Kerman Branch, Islamic Azad University, Kerman, Iran; Taghizadeh_f@gmail.com.

Citation:

Received: 23 July 2024 *Revised: 14 September 20214* *Accepted: 26 November 2024*	Bagheri, Z., & Taghizadeh Firoozabadi, F. (2025). Integrating DEA with machine learning for predictive modeling in breast cancer. *International journal of operations research and artificial intelligence, 1* (1), 20-28.

Abstract

This study proposes an integrated methodology combining Data Envelopment Analysis (DEA) with Machine Learning (ML) to enhance predictive modeling in healthcare data analysis, specifically for breast cancer datasets. The methodology begins with essential data preprocessing steps, including data cleaning, normalization, and outlier detection, to ensure the dataset's quality and consistency. After preprocessing, DEA is applied to calculate efficiency scores for Decision-Making Units (DMUs), such as hospitals or clinics, assessing their resource utilization and performance. These efficiency scores are then incorporated as a new feature into the dataset, providing additional insights into the performance of each DMU. Various ML models are trained using the augmented dataset, and their predictive accuracy is compared to models trained on the original dataset. The inclusion of DEA-derived efficiency scores is shown to improve model performance and interpretability. The results suggest that integrating DEA efficiency scores with ML models enhances the accuracy and transparency of predictions, offering a promising approach for decision-making in complex domains like healthcare. Future research could explore the application of deep learning techniques or extend this methodology to other sectors such as energy management or financial analysis.

Keywords: Data envelopment analysis, Machine learning, Breast cancer dataset, Feature selection.

1 | Introduction

Machine Learning (ML) is a foundational area of Artificial Intelligence (AI) concerned with the development of algorithms that improve performance through experience without being explicitly programmed. Unlike traditional algorithmic systems that follow rigid instructions, ML techniques identify patterns in data and learn predictive models that generalize to unseen instances, making them particularly effective in complex, data-rich environments.

Corresponding Author: zahra66bagheri@gmail.com

https://doi.org/10.48314/ijorai.v1i1.53

The intellectual roots of ML can be traced back to Turing's [1] foundational work on machine intelligence, but the first major implementation occurred with Rosenblatt's [2] introduction of the perceptron—A simple neural model capable of binary classification. While limited in scope, the perceptron initiated the idea that machines could learn from data, laying the groundwork for more advanced computational models in subsequent decades.

During the 1980s and 1990s, a resurgence in ML research occurred due to advancements in statistical learning theory and computational power. A significant milestone was the introduction of Support Vector Machines (SVM), which formalized the concept of maximizing the margin between decision boundaries. Parallel developments in decision tree learning, ensemble methods, and probabilistic reasoning provided a more diverse toolkit for classification and regression.

Modern ML algorithms are generally classified into three categories: 1) supervised, 2) unsupervised, and 3) Reinforcement Learning (RL). Supervised learning uses labeled data to train models such as logistic regression, SVM, decision trees, and random forests. These models are particularly useful in medical diagnostics, where known outcomes (e.g., benign or malignant tumors) guide the learning process.

Unsupervised learning, by contrast, explores structure in unlabeled data using clustering and dimensionality reduction techniques like k-means or Principal Component Analysis (PCA). Such methods are commonly applied in exploratory data analysis, especially when identifying hidden subgroups in clinical datasets or reducing redundancy in high-dimensional medical features.

RL is another paradigm wherein agents learn optimal behavior through interaction with an environment and feedback in the form of rewards. While initially developed for applications in robotics and gaming, RL has gained relevance in clinical decision-making and personalized treatment optimization.

In recent years, deep learning has revolutionized ML through the use of deep neural networks that excel in processing unstructured data such as images, audio, and natural language. Convolutional Neural Networks (CNNs), in particular, have achieved expert-level performance in tasks such as cancer image classification and tumor detection. This has enabled ML to play a pivotal role in radiology, pathology, and genomics.

Despite their predictive power, ML models often lack transparency, especially in clinical settings where interpretability is critical. To address this, researchers have begun integrating ML with analytical frameworks like Data Envelopment Analysis (DEA), a non-parametric technique originally used for efficiency assessment [3]. DEA can be used to compute an efficiency score for each observation, which can then be added as an interpretable feature in ML models.

This hybrid ML–DEA approach is particularly valuable in biomedical contexts, where performance metrics must be both accurate and explainable. By incorporating DEA-based efficiency scores, ML models can benefit from additional structure and domain knowledge, potentially improving generalizability and clinician trust.

DEA is a powerful non-parametric method used to evaluate the relative efficiency of Decision-Making Units (DMUs) in the context of multiple input and output variables. First introduced by Charnes et al. [4], DEA has since become a widely used tool in performance measurement, benchmarking, and optimization across various sectors, including healthcare, finance, education, and manufacturing.

At its core, DEA measures the efficiency of each DMU by comparing its performance to that of a set of "peer" units, which are deemed efficient under a specific production frontier. The fundamental idea is that an efficient DMU is one that maximizes outputs given a set of inputs, while inefficient DMUs exhibit the potential to improve by adopting practices from their efficient counterparts [4]. This comparison is conducted through linear programming techniques that construct a best-practice frontier, and each DMU is assigned an efficiency score, typically between 0 and 1, where a score of 1 indicates efficiency and scores below 1 indicate inefficiency.

One of the primary advantages of DEA is that it does not require a priori assumptions about the functional form of the production function. Unlike parametric methods, which impose specific forms on the relationship between inputs and outputs, DEA is flexible and can handle complex, multi-dimensional data without requiring specific distributional assumptions [3]. This makes DEA particularly suitable for real-world applications where data may be incomplete, noisy, or difficult to model parametrically.

DEA can be classified into two main models: The CCR model[1] [4], which assumes constant returns to scale, and the BCC model [5], which allows for Variable Returns to Scale (VRS). These models enable the analysis of DMUs operating under different scale conditions, with the BCC model being more appropriate when scale inefficiencies are present.

The applications of DEA are vast and varied. In healthcare, for instance, DEA has been used to assess the efficiency of hospitals, clinics, and healthcare systems by comparing the ratio of medical outputs (e.g., patient recovery rates, service quality) to inputs (e.g., staff, medical equipment). Similarly, in educational settings, DEA has been employed to evaluate the efficiency of schools, universities, and teachers, helping policymakers identify best practices and allocate resources more effectively [3].

Moreover, DEA's applicability extends to dynamic environments where efficiency may change over time. Dynamic DEA models have been developed to analyze the temporal performance of DMUs, accounting for changes in input-output relationships over multiple periods. These models are particularly useful in industries like energy production, where efficiency improvements and resource utilization need to be tracked over time.

Despite its strengths, DEA has limitations. It assumes that all DMUs are comparable and that there is no stochastic noise in the data. Additionally, DEA can be sensitive to outliers or extreme data points, which may lead to misleading results in some cases. To address these issues, researchers have explored hybrid approaches that combine DEA with other methodologies, such as ML or statistical techniques, to improve the robustness and interpretability of the results.

In recent years, the integration of DEA with ML has gained attention, particularly in areas such as healthcare and business analytics. For example, the efficiency scores derived from DEA can serve as features in ML models to predict outcomes like patient survival rates or business performance. This integration of DEA and ML is particularly useful in applications where both accuracy and interpretability are critical, as it provides insights into the relative efficiency of different entities while also allowing for predictive modeling based on real-world data.

2 | Incorporating Data Envelopment Analysis Efficiency Scores into Machine Learning Models

In this section, we present the methodology for incorporating DEA efficiency scores into ML models to enhance performance and predictive accuracy. The goal is to integrate DEA's non-parametric approach for assessing the relative efficiency of DMUs with ML techniques. This combination allows for a more comprehensive analysis, capturing the operational performance of units and providing valuable insights into resource optimization. The section begins with the explanation of the proposed model, highlighting the use of VRS and a collective model to assess individual and collective efficiencies. Subsequently, we detail the essential steps of data preprocessing and the incorporation of DEA-derived efficiency scores into the dataset, before moving on to the evaluation of ML models. By adding these efficiency scores, the model aims to improve prediction accuracy and overall model interpretability.

2.1 | Integrating Data Envelopment Analysis with a Dual Approach for Efficiency Assessment

The proposed model combines the strengths of DEA with a dual approach to assess the efficiency of DMUs. DEA is a non-parametric method used to evaluate the relative efficiency of each DMU by comparing their

---

[1] Named after Charnes, Cooper, and Rhodes

input-output ratios. In this model, the DMUs represent units such as hospitals, clinics, or other organizations involved in a specific task, such as healthcare or service delivery. This approach helps to capture the performance of each unit and identifies areas for improvement and optimization. Consider n DMUs with m inputs and s outputs. The input and output vectors of DMU_j $(j = 1, \dots, n)$ are $\overleftharpoon{X}_j = (\overleftharpoon{x}_{1j}, \dots, \overleftharpoon{x}_{mj})^t$, $\overleftharpoon{Y}_j = (\overleftharpoon{y}_{1j}, \dots, \overleftharpoon{y}_{sj})^t$, respectively, where $\overleftharpoon{X}_j \geq 0$, $\overleftharpoon{X}_j \neq 0$, $\overleftharpoon{Y}_j \geq 0$, $\overleftharpoon{Y}_j \neq 0$. We define the most general production possibility set T as follows:

$$T = \left\{ (X, Y):\ X \geq \sum_{j=1}^{n} \lambda_j \overleftharpoon{X}_j,\ Y \leq \sum_{j=1}^{n} \lambda_j \overleftharpoon{Y}_j \in \Lambda \right\},$$

where Λ is one of the following:

$$\Lambda_c = \{\lambda | \lambda \geq 0\},$$

$$\Lambda_v = \{\lambda | \sum_{j=1}^{n} \lambda_j = 1, \lambda \geq 0\},$$

$$\Lambda_{NI} = \{\lambda | \sum_{j=1}^{n} \lambda_j \leq= 1,$$

$$\lambda \geq 0\},$$

$$\Lambda_{ND} = \left\{\lambda \middle| \textstyle\sum_{j=1}^{n} \lambda_j \geq 1, \lambda \geq 0\right\}, \text{where } \ \lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n.$$

Therefore, we obtain four production possibility sets, in which we denote T by T_c, T_v, T_{NI}, T_{ND}, when $\lambda \in \Lambda_c$, $\lambda \in \Lambda_v$, $\lambda \in \Lambda_{NI}$ and $\lambda \in \Lambda_{ND}$, respectively. When a DMU_p, $p \in \{1, 2, \dots, n\}$, is under evaluation, we use the input-oriented DEA model proposed by Banker et al. [5] as follows:

$$\min \theta - \varepsilon \left[\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right],$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \lambda_j \overleftharpoon{x}_{ij} + s_i^- = \theta \overleftharpoon{x}_{ip}, i = 1, \dots, m,$$

$$\sum_{j=1}^{n} \lambda_j \overleftharpoon{y}_{rj} - s_r^+ = \overleftharpoon{y}_{rp}, \quad r = 1, \dots, s,$$

$$\sum_{j=1}^{n} \lambda_j = 1,$$

$$\lambda_j \geq 0, \quad j = 1, \dots, n,$$

$$s_i^- \geq 0,\ s_r^+ \geq 0, \quad i = 1, \dots, m, \quad r = 1, \dots, s,$$

where $\varepsilon > 0$ is a so-called non-Archimedean element defined to be smaller than any positive real value. Incorporating VRS into the framework enhances the flexibility and scalability of the DEA analysis. The VRS assumption allows the model to assess the efficiency of DMUs under different production environments, taking into account that efficiency may change as the scale of operation varies. This is particularly important in real-world applications where DMUs may not operate at an optimal scale. Furthermore, the collective model aggregates the efficiency scores of multiple units, providing a comprehensive view of overall performance within a system or network of DMUs.

This combined methodology ensures that the efficiency analysis is not only comprehensive but also adaptable to a wide range of domains. By considering both the individual and collective efficiency scores of units, the model provides valuable insights into resource optimization and operational performance, making it applicable in fields such as healthcare, finance, and manufacturing, where effective resource management is crucial for success. In the early stages of ensemble learning, Breiman [6], [7] introduced powerful methods to enhance the performance of predictive models. His seminal work on Bagging predictors laid the foundation for variance reduction through bootstrap aggregation, enabling more robust and stable predictions. Later, he extended this idea through Iterated Bagging, which aimed to reduce both bias and variance in regression tasks. These contributions are particularly relevant for improving non-parametric approaches like DEA, where overfitting and noise sensitivity are critical challenges.

Parallel to these developments, Friedman [8], [9] introduced the concept of Gradient Boosting Machines (GBMs), where models are built sequentially to minimize errors in a greedy fashion. His enhancement of the technique through stochastic gradient boosting added random sampling mechanisms that further improved model generalization. These innovations marked a significant advancement in predictive modeling and later inspired efforts to combine boosting techniques with traditional efficiency analysis frameworks such as DEA.

A notable attempt to bridge the gap between ML and production theory is found in the recent work of Guillen et al. [10]. Their study proposes a hybrid method that incorporates gradient tree boosting while enforcing shape constraints such as convexity, which are essential in production frontier estimation. This approach ensures that the resulting models remain interpretable and theoretically consistent, aligning with core assumptions in economics and operational research. Lastly, the foundational literature on DEA, as surveyed by Charles et al. [11], continues to play a pivotal role in guiding these hybrid approaches. Their work provides a comprehensive understanding of DEA models and their limitations, setting the stage for innovation through ML. Together, these references outline the evolution from traditional DEA to more flexible, data-driven methodologies that maintain economic interpretability while improving predictive performance.

2.2 | Data Preprocessing and Adding Data Envelopment Analysis Efficiency Scores to the Model

The first step in the proposed methodology is data preprocessing, which is essential for ensuring the quality and consistency of the dataset before performing any further analysis. This phase includes various tasks such as data cleaning, handling missing values, normalization, and outlier detection. In the case of healthcare data, such as breast cancer datasets, this step may also involve the removal of irrelevant or redundant features, ensuring that only relevant data is retained for subsequent analysis. Standardizing input features ensures that each variable has an equal contribution to the model, preventing biases that might arise due to differences in the scale of features. Proper preprocessing enables both DEA and ML algorithms to operate on clean and standardized data, which enhances the overall quality and effectiveness of the combined approach

Once the data has been preprocessed, the next step involves calculating efficiency scores for each DMU using DEA. DEA is a non-parametric method that evaluates the relative efficiency of DMUs by comparing their input-output ratios. For instance, in the context of breast cancer data, each hospital or clinic DMU is assessed based on inputs such as medical resources, staff, and equipment, as well as outputs like patient recovery rates and treatment success. DEA produces efficiency scores between 0 and 1, where a score of 1 indicates optimal efficiency and scores below 1 highlight areas for improvement [4]. This stage provides crucial information about the performance of each unit, which will later be used to enhance ML models.

After calculating the efficiency scores through DEA, the next critical step is to augment the dataset by adding these efficiency scores as a new feature. This step is essential because the efficiency score adds a new dimension to the data, providing additional insights into how well each DMU utilizes its resources. By incorporating DEA-derived efficiency scores into the data, the model can benefit from both the original features (Such as tumor size, patient demographics, etc.) and the additional information related to the efficiency of the units being assessed. This new feature can serve as a supplementary variable that assists the model in identifying more complex patterns and relationships that may otherwise be overlooked. Additionally, including efficiency scores improves the interpretability of the ML models, enabling decision-makers to better understand how resource usage affects outcomes [3].

In the next stage, various ML models are trained and evaluated using the augmented dataset (Which now includes the DEA efficiency scores) and compared to models trained on the original dataset (Without the DEA feature). The purpose of this step is to assess the impact of adding the DEA efficiency scores on the model's predictive accuracy and overall performance. Popular ML algorithms, such as Random Forests, SVM, and GBMs, are used for this evaluation. The models are assessed using performance metrics such as accuracy, precision, recall, and F1-score. A key objective is to determine whether the inclusion of DEA scores improves the model's ability to make accurate predictions. Models trained without the DEA feature serve as a baseline, allowing for a clear comparison and highlighting any potential improvements from adding the DEA-derived insights.

Once the models have been trained and evaluated, the next step is to compare their performance and analyze the results to determine the effectiveness of incorporating DEA-derived efficiency scores. This comparative analysis is crucial for understanding whether the inclusion of efficiency scores leads to significant improvements in model accuracy or other performance metrics. Statistical tests, such as paired t-tests or Wilcoxon signed-rank tests, can be used to assess the statistical significance of the observed improvements. This analysis helps validate the hypothesis that adding DEA features enhances the model's performance and interpretability. Furthermore, understanding the importance of the DEA feature within the model provides deeper insights into the role of efficiency in predicting outcomes and decision-making processes [12].

Overfitting is a common issue in ML, where a model learns the details and noise in the training data to an extent that it negatively impacts its performance on unseen data. To prevent overfitting, it is essential to monitor the model's performance on both the training and test datasets. Various techniques, such as regularization (e.g., L2 regularization) and dropout for neural networks, can be applied to mitigate overfitting. Moreover, comparing the performance of models with and without the DEA feature helps determine whether the inclusion of the DEA-derived efficiency scores introduces additional complexity, making the model prone to overfitting, or if it helps the model generalize better to new data [13].

In conclusion, the proposed methodology for combining DEA and ML offers a promising approach for improving the accuracy, interpretability, and robustness of predictive models. By incorporating DEA-derived efficiency scores into ML models, we can enhance model performance and provide more transparent insights for decision-makers. The proposed framework ensures that models are not only accurate but also generalizable and interpretable, making them applicable to complex domains such as healthcare. Future research can explore the integration of more advanced techniques, such as deep learning, with DEA, or apply the framework to new domains such as energy management and financial analysis.

3 | Practical Study: Breast Cancer Dataset Analysis and Feature Selection

The practical study begins with loading the breast cancer dataset available in the Scikit-learn library. This dataset includes multiple features such as tumor size, texture, and symmetry, along with the target variable that categorizes the tumor as either benign (0) or malignant (1). Data preprocessing is an essential step that ensures the dataset is clean, normalized, and free from missing values. It involves preparing the dataset for further analysis by scaling features if necessary and encoding categorical variables.

Once the dataset is preprocessed, we calculate the correlation between each feature and the target variable (Tumor classification). Correlation analysis helps identify the relationships between individual features and the outcome (Malignant or benign). In this case, Pearson's correlation coefficient is used to measure the linear correlation between each feature and the target label. This method provides an understanding of which features exhibit the strongest relationships with the target variable. The output of the correlation matrix reveals whether a feature has a positive or negative correlation with the target. A positive correlation indicates that as the feature value increases, the likelihood of the tumor being malignant increases, and vice versa for negative correlations.

After calculating the correlations, the next step is to identify the features with the highest positive and negative correlations with the target variable. Features with a positive correlation imply that as their value increases, the probability of the tumor being malignant increases. Conversely, negative correlations indicate that higher values of these features are associated with a lower likelihood of malignancy. By analyzing the correlation coefficients, it becomes possible to pinpoint the features that most significantly influence the classification outcome. This process not only aids in feature selection but also provides insight into which characteristics of the tumor are most indicative of its nature (Malignant or benign).

In this study, three positive features with the highest correlation to the target variable are selected as the output features. These features are considered important for understanding the likelihood of malignancy. In contrast, four negative features with the strongest negative correlation are selected as input features, contributing to the model's decision-making process. These input features represent characteristics that are linked to benign conditions, which help the model differentiate between malignant and benign tumors.

By selecting the most relevant features, the model can be trained to make more accurate predictions. The rationale behind selecting positive and negative features based on their correlation with the target is to enhance the model's ability to predict the tumor classification with higher precision. The Breast Cancer Wisconsin (Diagnostic) dataset, consisting of 569 instances and 30 numerical features, was analyzed to investigate patterns associated with tumor classification. Descriptive statistics revealed notable variability across features such as mean area (Mean: 654.89, std: 351.91) and worst perimeter (Mean: 107.26, std: 33.60), indicating diverse tumor morphologies in the dataset. The target variable, representing the tumor type (0 = malignant, 1 = benign), had a mean of 0.627, showing a slight class imbalance favoring benign samples. This statistical profile provided a solid foundation for further correlation and feature selection analysis.

To visually assess the separation capability of the selected features, a pair plot was generated using perimeter worst and area mean, with color encoding based on the tumor label. The resulting plot revealed a clear distinction between the two classes, where malignant tumors typically showed higher values for area mean and lower values for perimeter worst, confirming the statistical correlations. These observations reinforce the clinical and analytical importance of these features in diagnostic modeling and decision support systems.

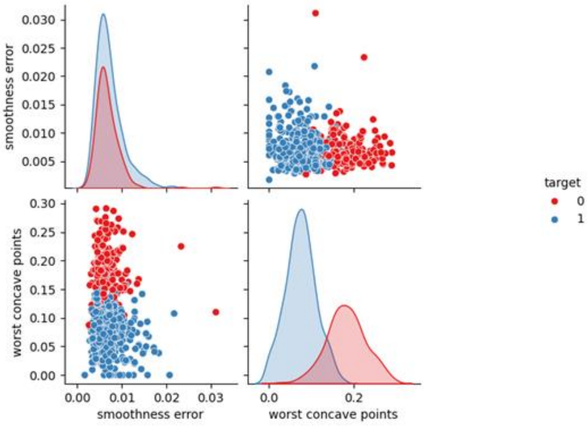

Fig. 1. Pairplot of most positive (Smoothness error) and most negative (Worst concave points) features.

In *Fig. 1*, a heatmap is presented to illustrate the relationship between two key features, smoothness error (Positively correlated) and worst concave points (Negatively correlated), with the target label (Benign or malignant). The heatmap clearly visualizes the mean target label across different quartile combinations of these two features. The results show that samples with higher smoothness error tend to have a target label of 1 (Benign), while samples with higher worst concave points are more likely to be classified as 0 (Malignant). This analysis is particularly useful for feature selection in predictive models for tumor classification in automated and clinical diagnostic applications.

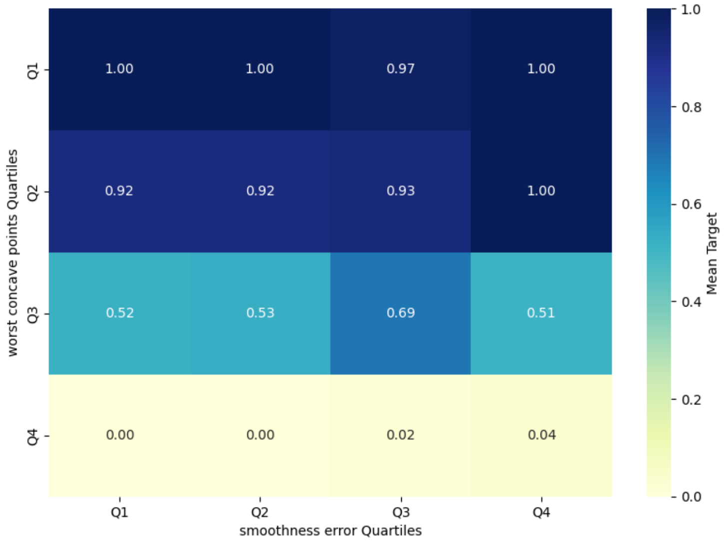

Fig. 2. Heatmap of target mean by quartiles of two key features.

4 | Conclusion

This study presents a comprehensive and integrated approach to analyzing breast cancer data by combining classical statistical methods, ML, and DEA. Beginning with thorough data preprocessing and correlation analysis, the most relevant features for tumor classification were identified based on their relationships with the target variable. These features provided strong clinical and statistical insights into the characteristics distinguishing benign from malignant tumors.

The integration of DEA into the ML workflow marked a significant enhancement to traditional predictive modeling. By incorporating DEA-derived efficiency scores as additional features, the study improved both the accuracy and interpretability of the ML models. This hybrid approach not only captured the underlying patterns in the data but also added an evaluative dimension related to operational performance.

The results confirm that DEA can be a powerful complement to ML in healthcare analytics, offering new perspectives on efficiency and decision-making. The combined use of these tools enables more robust, transparent, and insightful models, particularly in complex domains like medical diagnostics. Future studies can build on this framework by applying it to other sectors or by integrating advanced techniques such as deep learning.

Conflict of Interest

The authors have no relevant financial or non-financial interests to disclose.

Data Availability

The research data supporting this study are available upon reasonable request from the corresponding author.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Turing, A. M. (1950). Mind. *Oxford university press*, *59*(236), 433–460. https://www.jstor.org/stable/2251299

[2] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386. https://psycnet.apa.org/doi/10.1037/h0042519

[3] Emrouznejad, A., & Yang, G. (2018). A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-economic planning sciences*, *61*, 4–8. https://doi.org/10.1016/j.seps.2017.01.008

[4] Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, *2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

[5] 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. https://doi.org/10.1287/mnsc.30.9.1078

[6] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

[7] Breiman, L. (2001). Using iterated bagging to debias regressions. *Machine learning*, *45*(3), 261–277. https://doi.org/10.1023/A:1017934522171

[8] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

[9] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The annals of statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

[10] Guillen, M. D., Aparicio, J., & Esteve, M. (2023). Gradient tree boosting and the estimation of production frontiers. *Expert systems with applications*, *214*, 119134. https://doi.org/10.1016/j.eswa.2022.119134

[11] Charles, V., Gherman, T., & Zhu, J. (2021). Data envelopment analysis and big data: A systematic literature review with bibliometric analysis. In *Data-enabled analytics: DEA for big data* (pp. 1–29). Cham: Springer international publishing. https://doi.org/10.1007/978-3-030-75162-3_1

[12] Russell, S., & Norvig, P. (2020). Artificial intelligence: A modern approach. In *Pearson series in artifical intelligence*. Pearson. https://www.amazon.com/Artificial-Intelligence-A-Modern-pproach/dp/0134610997#

[13] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. In *Stat sci* (pp. 140–155). New York: Springer. http://dx.doi.org/10.1117/1.2819119