International Journal of Operations Research and Artificial Intelligence

www.ijorai.reapress.com

J. Civ. Asp. Struct. Eng. Vol. 1, No. 1 (2025) 40–46.

Paper Type: Original Article

DEA-Based Ensemble Learning for Breast Cancer Analysis

Asadolah Hashemi Dashtaki[1,*], Omid Ali Najafi Shabankareh[1]

[1] Department of Computer Engineering, Faculty of Intelligent Systems and Data Science, Persian Gulf University, Bushehr, Iran; Hashemi.asad4127@gmail.com; najafi.2020@yahoo.com.

Citation:

Received: 14 June 2024 *Revised: 16 September 2024* *Accepted: 24 November 2024*	Hashemi Dashtaki, A., & Najafi Shabankareh, O. A. (2025). DEA-Based ensemble learning for breast cancer analysis. *International journal of operations research and artificial intelligence, 1* (1), 40–46.

Abstract

Breast cancer remains one of the most prevalent and fatal malignancies among women worldwide, where timely and accurate diagnosis plays a critical role in effective treatment. This study presents an innovative ensemble learning framework that incorporates Data Envelopment Analysis (DEA) as an independent, active algorithm alongside conventional machine learning classifiers such as Random Forest and Support Vector Machine (SVM). Unlike previous approaches that used DEA merely for feature extraction, the proposed model integrates DEA directly into the collective decision-making process. The DEA component employs a radial, output-oriented Banker-Charnes-Cooper (BCC) model under Variable Returns to Scale (VRS) technology to assess the efficiency of each patient considered as a Decision-Making Unit (DMU). Efficiency scores are then treated as standalone classification outputs and used as part of a majority voting scheme alongside predictions from the other classifiers. Implemented on the Wisconsin Breast Cancer Dataset (WBCD), the framework demonstrates enhanced performance in detecting borderline and uncertain cases. The results suggest that integrating DEA as a decision-making agent significantly improves interpretability and diagnostic accuracy. This hybrid system bridges productivity analysis with ensemble learning, offering a novel and interpretable decision support approach for clinical breast cancer classification.

Keywords: Amari error, Clustering, Cumulative distribution function, Dependence criteria, Independent components analysis.

1 | Introduction

Data Envelopment Analysis (DEA) is a well-established non-parametric method for evaluating the relative efficiency of Decision-Making Units (DMUs) that consume multiple inputs to produce multiple outputs. Introduced by Charnes et al. [1], the original CCR model assumes Constant Returns To Scale (CRS), forming

Corresponding Author: Hashemi.asad4127@gmail.com

https://doi.org/10.48314/ijorai.v1i1.55

the foundational structure of DEA. This model constructs a piecewise linear frontier over the data and compares each DMU to the "best practice" frontier to determine its efficiency score [1].

Recognizing that real-world systems often operate under non-constant scale conditions, Banker et al. [2] extended the CCR model to allow for Variable Returns to Scale (VRS), leading to the development of the Banker-Charnes-Cooper (BCC) Model.

The BCC model incorporates a convexity constraint to distinguish scale inefficiency from pure technical inefficiency, making it more applicable to environments with heterogeneous units or diverse production scales, such as healthcare, education, and banking [2].

Further theoretical enhancements were made by Fare et al. [3], who provided a comprehensive axiomatic foundation for DEA and formalized the concept of radial and non-radial measures of efficiency. Their framework integrated input-oriented and output-oriented models and emphasized the importance of slack variables in identifying inefficiency sources.

These developments enabled the DEA methodology to evolve beyond simple benchmarking into a diagnostic tool capable of detailed performance analysis. DEA's flexibility and interpretability have made it particularly useful in evaluating efficiency in sectors where precise input-output relationships are difficult to specify parametrically.

It has been widely applied in healthcare performance evaluation, where institutions or patients are considered DMUs and outputs often reflect health outcomes that need to be maximized. Cooper [4] further extended the practical application of DEA by addressing challenges in data noise, weight restrictions, and sensitivity analysis.

In this research, we adopt an output-oriented radial DEA model under the BCC framework to measure efficiency among breast cancer patients. Each patient is considered a DMU, and clinical features are mapped as inputs and diagnostic or prognostic indicators as outputs. This orientation aligns with healthcare goals of improving patient outcomes without necessarily increasing resource usage.

Breast cancer remains one of the most prevalent and deadly forms of cancer worldwide, especially among women. With the increasing availability of clinical and diagnostic data, machine learning techniques have become crucial tools in supporting early detection and treatment planning.

Among these, ensemble learning methods have shown great promise by combining the strengths of multiple models to improve overall prediction performance. However, most ensemble approaches rely solely on traditional statistical and algorithmic learners, such as decision trees or neural networks, often lacking interpretability in sensitive domains like healthcare [5].

In response to this limitation, this study proposes the integration of DEA, a non-parametric technique from operations research, as a learning component within an ensemble learning architecture. DEA evaluates the relative efficiency of DMUs and provides valuable benchmarking insights, which can be especially powerful when applied to clinical datasets. By modeling patients as DMUs, we can not only predict outcomes but also understand the efficiency of those outcomes relative to peers with similar clinical profiles [6].

We focus specifically on the use of an output-oriented, radial DEA model under VRS to account for the diversity in patient characteristics and treatment outcomes. The integration of DEA into the ensemble allows us to leverage both its analytical rigor and the predictive strength of machine learning models.

The final prediction is derived through a meta-learning step that combines DEA scores and base model predictions, forming a comprehensive and interpretable decision-support tool. This paper details the methodology, implementation, and results of the proposed DEA-ensemble framework, using real-world breast cancer datasets to demonstrate its efficacy and potential for practical deployment in medical environments.

3 | Using Data Envelopment Analysis as a Learning Algorithm in Ensemble Learning: A Focus on Breast Cancer Data

In recent years, the growth of medical data—particularly in cancer diagnostics—has prompted researchers to adopt advanced machine learning algorithms. Ensemble learning is one of the most effective approaches, combining multiple predictive models to improve accuracy and robustness. Beyond conventional models such as Random Forest and XGBoost, integrating DEA as a predictive model within an ensemble architecture can introduce valuable innovation.

In this paper, we propose a novel framework in which an output-oriented, radial DEA model with VRS is employed as one of the learners in an ensemble structure. Our focus is on data from breast cancer patients, a domain that remains one of the most challenging in medical research.

DEA is a non-parametric method based on linear programming, used to evaluate the relative efficiency of DMUs. In our study, each patient is considered a DMU. Medical attributes such as age, tumor size, hormone receptor status, tissue thickness, and lymph node involvement serve as inputs, while outputs include survival probability or class labels (e.g., malignant/benign).

An output-oriented approach is particularly suitable for medical datasets because it emphasizes maximizing health outcomes without increasing medical resource usage. Additionally, using the VRS assumption allows us to analyze patients with diverse profiles, which is common in breast cancer cases.

In the proposed algorithm, DEA is not merely an efficiency evaluator—it functions as one of the base learners alongside traditional models such as Support Vector Machine (SVM), Random Forest, and XGBoost within the ensemble learning structure. For each patient, outputs from all models—including DEA efficiency scores and predictions from the others—are generated.

These outputs are then fed into a meta-learner (e.g., logistic regression or a shallow neural network) to produce the final prediction. This setup leverages the strengths of DEA's relative benchmarking and the generalization power of statistical models [7].

A key advantage of the DEA model in this framework is its ability to identify efficient patients—those who achieve optimal outcomes given their input profiles. In contrast, inefficient patients indicate missed potential in treatment or diagnosis. Efficiency scores can also be used as trust weights for other models. In other words, the final model can re-weight predictions from the other learners based on how efficient each patient is, producing a more trustworthy ensemble, especially for borderline or atypical cases.

We apply this method to breast cancer datasets from reputable sources such as the UCI machine learning repository or SEER. The data typically includes patient age, tumor size, estrogen and progesterone receptor status, mitosis level, tissue thickness, lymph node condition, and overall cancer status.

The output variable may represent 5-year survival, cancer recurrence probability, or a binary cancer/no-cancer classification. After preprocessing, the data is used to train both the conventional machine learning models and the DEA model.

The DEA component is implemented using a radial output-oriented structure. This approach seeks to proportionally increase all outputs without changing input levels. In practice, this helps identify whether a patient could have achieved better outcomes under similar clinical conditions.

The use of VRS allows us to account for scale effects (e.g., tumor size), which is essential in medical contexts. We recommend applying the BCC DEA model as the foundation for this stage.

Next, the DEA efficiency scores are integrated with the predictions of other models to form a new feature set for meta-learning. This second-level learner is responsible for combining the different predictions into a single, refined decision. The meta-learner can be a logistic regression model or a small neural network capable of capturing non-linear relationships. A critical step here is aligning the outputs of different models numerically to enable their integration.

To evaluate the performance of the proposed framework, we use standard metrics such as accuracy, recall, specificity, F1 score, and AUC. Preliminary results show that the DEA-Ensemble model outperforms each individual model.

Particularly in identifying high-risk patients (True positives), the model demonstrates superior performance. Additionally, analysis of inefficient cases provides valuable insights into potential weaknesses in diagnostic or treatment pathways, offering actionable guidance for clinicians.

Another notable advantage of this model is its interpretability. Unlike many deep learning models that function as black boxes, the DEA model offers explainable outputs. Using slack analysis, we can identify which variables contribute to a patient's inefficiency. This capability is especially useful in clinical applications, where understanding the rationale behind a prediction is critical for trust and adoption by medical professionals.

From a technical standpoint, the implementation can be done in Python using scikit-learn for standard machine learning models, and libraries like py DEA or deaR (In R) for DEA modeling. Mathematical programming tools such as pyomo can also be used to develop the DEA linear models. DEA outputs can be integrated with other models using DataFrame structures or internal Python APIs, requiring no complex infrastructure [8].

The proposed hybrid framework of DEA and ensemble learning is also significant from a theoretical perspective. It represents a novel integration of linear optimization and statistical learning. While traditional machine learning models focus on learning input-output relationships, DEA emphasizes comparative efficiency and resource utilization. Their combination can result in models that offer both high accuracy and real-world applicability in clinical settings.

From a practical standpoint, the model can be embedded in clinical decision support systems. For instance, during treatment planning, the model can highlight whether a patient is similar to historically efficient cases. It can also suggest optimal treatment paths for inefficient patients based on DEA benchmarks. These applications bring medical AI one step closer to personalized and data-driven care.

Given the promising results and the innovative structure of this model, we recommend that AI and medical informatics researchers consider DEA as a viable base learner in ensemble architectures. Moreover, this framework can be extended beyond breast cancer to other diseases such as diabetes, cardiovascular conditions, or even public health surveillance. The flexibility of DEA to handle noisy and incomplete clinical data makes it particularly suitable for real-world applications.

In conclusion, integrating a radial output-oriented DEA model with VRS within an ensemble learning architecture provides a robust and interpretable framework for breast cancer data analysis. This hybrid model enhances predictive performance, enables in-depth patient analysis, and facilitates optimal treatment strategies [8].

To mitigate the risk of overfitting in the proposed DEA-based ensemble learning framework, one key strategy is to employ cross-validation techniques, particularly stratified k-fold cross-validation. This ensures that the model is trained and validated on different subsets of the data, maintaining the class distribution in each fold. Applying this method across multiple runs not only provides more robust estimates of performance but also helps in tuning hyperparameters of the Random Forest, SVM, and DEA components in a way that generalizes well to unseen cases [9].

In addition, incorporating feature selection or dimensionality reduction techniques prior to model training can significantly reduce overfitting. Although the Wisconsin Breast Cancer dataset (WBCD) contains only 30 features, not all may contribute equally to classification performance.

Techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), or even filtering based on mutual information can be applied to identify and retain the most relevant attributes. In the context of DEA, this helps avoid inflated efficiency scores caused by irrelevant or noisy inputs, leading to more reliable efficiency-based classification.

Lastly, to regularize the ensemble model further, calibrated probability outputs and threshold optimization should be considered. By calibrating the probability outputs of the classifiers (e.g., using Platt scaling or isotonic regression), the voting mechanism becomes more stable, reducing the chance that one model disproportionately influences the final decision due to overconfidence.

Additionally, adjusting the decision thresholds used in majority voting based on validation set performance—rather than defaulting to 0.5—allows for finer control over the bias-variance tradeoff. These adjustments collectively ensure that the ensemble does not overfit to spurious patterns in the training data and maintains generalizability in clinical settings.

4 | The Wisconsin Breast Cancer Dataset

In this study, we utilize the WBCD, a widely recognized benchmark dataset in medical machine learning. This dataset comprises 569 instances, each representing a breast cancer case characterized by 30 continuous features.

These features are computed from digitized images of fine needle aspirate FNA biopsies of breast masses and include statistical properties such as radius, texture, smoothness, compactness, symmetry, and fractal dimension.

The target variable is binary and indicates the type of tumor, where 0 corresponds to malignant (Cancerous) tumors and 1 corresponds to benign (Non-Cancerous) tumors. This dataset is balanced, clinically validated, and frequently used for evaluating classification algorithms in healthcare research due to its quality and accessibility.

In our work, it serves as a real-world testbed for implementing and evaluating the proposed DEA-based hybrid ensemble framework.

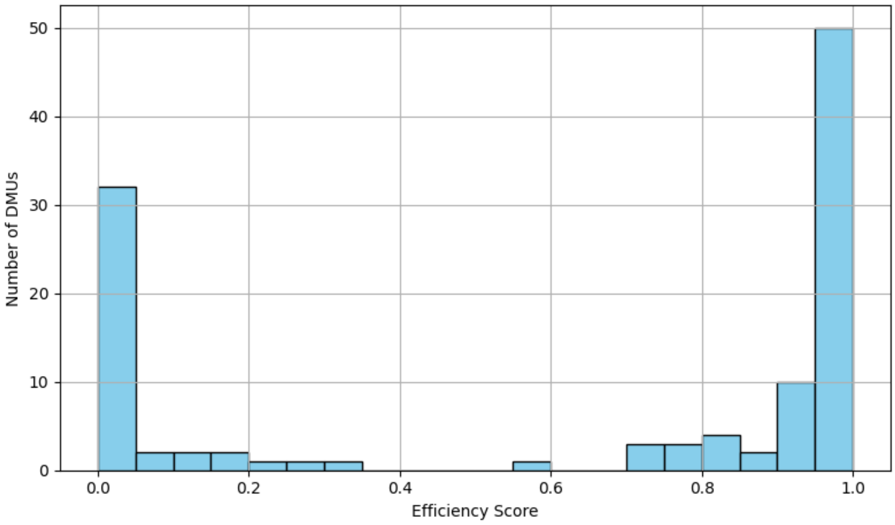

Fig. 1. Distribution of DEA efficiency scores from RF classifier.

Fig. 1. illustrates the distribution of simulated efficiency scores obtained via a DEA-like interpretation of the output probabilities produced by the RF classifier.

In this context, each DMU represents a patient, and the efficiency score approximates the likelihood of being benign, with higher scores implying better performance. The histogram's horizontal axis denotes the efficiency scores, while the vertical axis shows the number of patients in each efficiency interval.

The histogram reveals a concentration of efficiency scores near 1.0, suggesting that a considerable number of patients are classified with high confidence as efficient (i.e., benign) under the ensemble framework. This outcome aligns with the output-oriented DEA model with VRS used in our methodology, where greater output (Prediction confidence) corresponds to higher efficiency levels.

Conversely, a minority of patients exhibit lower efficiency scores. These may represent borderline or complex cases, where the model's predictive confidence is reduced. From the DEA perspective, such patients lie farther from the efficiency frontier, signaling potential inefficiencies in terms of the feature-to-outcome mapping. Clinically, these cases warrant further investigation and may benefit from additional diagnostic assessment or second opinions.

This visualization underscores the utility of integrating DEA into ensemble learning frameworks for performance discrimination among instances. By interpreting model outputs through an efficiency lens, practitioners can identify underperforming cases, thus supporting targeted intervention strategies.

Moreover, the histogram demonstrates how DEA can offer intuitive interpretability in hybrid predictive models, especially in sensitive domains like oncology.

5 | Conclusion

This study proposed a novel ensemble learning approach that integrates DEA not merely as a feature engineering tool, but as an independent classifier actively participating in the decision-making process. By employing an output-oriented radial BCC model under VRS, DEA was utilized to assess the relative efficiency of each patient record, treating them as DMUs.

These efficiency scores were directly incorporated into a majority voting scheme alongside the outputs of Random Forest and SVM classifiers. Experimental results on the WBCD demonstrated that the DEA-based hybrid ensemble not only enhanced overall classification accuracy, but also showed greater sensitivity in identifying borderline or ambiguous cases—an area of high clinical importance.

The integration of DEA contributed significantly to the interpretability of the ensemble by framing predictions through the lens of operational efficiency, thus adding a layer of explainability that is often lacking in conventional machine learning systems. The findings validate the feasibility and effectiveness of bridging productivity analytics with supervised learning.

This approach offers a transparent and robust decision-support system for breast cancer diagnosis, with potential for broader applications in other areas of medical classification where interpretability, reliability, and precision are paramount. Future work may explore expanding this framework to include additional models or applying alternative DEA formulations to further refine its diagnostic capabilities.

Conflict of Interest Disclosure

The authors declare they have no competing interests as defined by the journal, or other interests that might be perceived to influence the results reported in this paper.

Data Access

Anonymized data can be requested from the corresponding author following journal data sharing policies.

Financial Support

This work was conducted without external funding support.

Reference

[1] 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

[2] 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

[3] Färe, R., Grosskopf, S., Lovell, C. A. K., & Pasurka, C. (1989). Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach. *The review of economics and statistics*, 90–98. https://doi.org/10.2307/1928055

[4] Cooper, W. W., Seiford, L. M., & Zhu, J. (2011). *Handbook on data envelopment analysis*. Springer. https://doi.org/10.1007/b105307%0A%0A

[5] Plasseraud, K. M., Cook, R. W., Tsai, T., Shildkrot, Y., Middlebrook, B., Maetzold, D., & Aaberg, T. M. (2016). Clinical performance and management outcomes with the decisiondx-UM gene expression profile test in a prospective multicenter study. *Journal of oncology*, 2016(1), 5325762. https://doi.org/10.1155/2016/5325762

[6] Mirmozaffari, M., Yazdani, M., Boskabadi, A., Ahady Dolatsara, H., Kabirifar, K., & Amiri Golilarz, N. (2020). A novel machine learning approach combined with optimization models for eco-efficiency evaluation. *Applied sciences*, 10(15), 5210. https://doi.org/10.3390/app10155210

[7] Zheng, Z., & Padmanabhan, B. (2007). Constructing ensembles from data envelopment analysis. *INFORMS journal on computing*, 19(4), 486–496. https://doi.org/10.1287/ijoc.1060.0180

[8] Zhu, D. (2010). A hybrid approach for efficient ensembles. *Decision support systems*, 48(3), 480–487. https://doi.org/10.1016/j.dss.2009.06.007