International Journal of Operations Research and Artificial Intelligence

www.ijorai.reapress.com

J. Civ. Asp. Struct. Eng. Vol. 1, No. 1 (2025) 29–39.

Paper Type: Original Article

Using Modified SBM and Decision Tree to Find Targets in Breast Cancer Data

Sara Fanati Rashidi[1,*], Maryam Olfati[2]

[1] Department of Mathematics, Shiraz Branch, Islamic Azad University, Shiraz, Iran; sarafanati@yahoo.com.

[2] Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Czech Republic; maryamolfati61@gmail.com.

Citation:

Received: 4 October 2024 *Revised: 14 December 2024* *Accepted: 16 Februrary 2024*	Fanati Rashidi, S., & Olfati, M. (2025). Using modified SBM and decision tree to find targets in breast cancer data. *International journal of operations research and artificial intelligence, 1* (1), 29-39.

Abstract

This study presents a combined approach using a modified Slacks-Based Measure (SBM) and a Decision Tree algorithm to identify target patients within breast cancer data. Traditional Data Envelopment Analysis (DEA) models often classify multiple patients as equally efficient, which limits the ability to distinguish between them. By modifying the SBM model to better handle input and output slacks, we aim to capture more accurate efficiency levels. We apply this method to the Scikit-learn breast cancer dataset, treating each patient as a Decision-Making Unit (DMU). The Decision Tree algorithm is used to identify the most significant features influencing efficiency. These key features are assigned higher weights in the SBM model to refine the analysis. The results allow for the identification of biologically significant target patients who demonstrate distinct efficiency profiles. This approach offers a useful tool for discovering hidden patterns in medical data and supports data-driven decision-making in cancer diagnosis and treatment planning.

Keywords: Modified slacks-based measure model, Decision tree, Data envelopment analysis, Decision-making unit, Efficiency in cancer patients.

1 | Introduction

Data Envelopment Analysis (DEA) is a well-established non-parametric approach used to evaluate the relative efficiency of Decision-Making Units (DMUs) across various fields, including medicine, economics, and management. The four cited references are among the most foundational and widely cited works in the field of DEA and have played a significant role in the theoretical and practical development of this method. The classic paper by Charnes et al. [1] first introduced the CCR model, which, under the assumption of constant

Corresponding Author: sarafanati@yahoo.com

https://doi.org/10.48314/ijorai.v1i1.54

returns to scale, formally established DEA. Subsequently, Banker et al. [2] developed the BCC model, which considers Variable Returns to Scale (VRS) and helps more precisely analyze technical and scale inefficiencies. Later, Tone [3] enriched DEA by introducing the Slacks-Based Measure (SBM) model, which focuses on input and output slacks and presents a non-radial, fractional programming approach that is often more appropriate for real-world problems.

Finally, Emrouznejad and Yang [4] provided a comprehensive survey of 40 years of scholarly literature in DEA, documenting the evolution, major application areas, and future challenges of the field, and thus offering a key reference for researchers.

In most DEA models, efficient units receive a full efficiency score of one (Or 100%). However, in practice, multiple DMUs often achieve this "efficient status" simultaneously, making the discrimination among fully efficient units-known as the super-efficiency problem-a critical topic of research [5]. To address this, Tone [5] introduced the SBM, which evaluates efficiency by directly considering the slacks in inputs and outputs, unlike traditional radial models such as the Andersen and Petersen [6] method that overlook these slack variables.

Inspired by this foundation, the present study aims to apply a modified SBM model to assess the biological efficiency of patients using the breast cancer dataset provided in the Scikit-learn library. Rather than focusing solely on classification tasks, this research adopts a structural perspective to explore patterns of efficiency and inefficiency at the cellular level.

First, the original SBM model will be thoroughly introduced and its underlying logic explained. Next, the numerical and biological characteristics of the breast cancer dataset will be examined to prepare the groundwork for modeling. Subsequently, a modified SBM model will be implemented, allowing for more precise identification of inefficiencies and performance across individual patients based on their cellular features.

In the final stage, suitable input and output variables will be defined, and patients will be treated as DMUs for conducting a biological pattern recognition analysis based on the output of the modified SBM model. Additionally, to enhance the precision of the analysis and emphasize the most influential features, a Decision Tree algorithm will be used.

This algorithm will help identify which cellular attributes play the most significant roles in determining efficiency and will allow those features to be weighted more heavily within the SBM model, thus impacting the overall benchmarking and pattern recognition results.

This innovative integration of the SBM model, breast cancer biomedical data, and machine learning techniques provides a new framework for evaluating biological performance and ultimately leads to the identification of reference patients and more effective treatment pathways.

2 | Overview of Data Envelopment Analysis, Slacks-Based Measure, and Breast Cancer Data

The SBM model, which forms the core of our efficiency analysis, is presented in Section 1. Section 2 introduces and explains the breast cancer dataset used in this study. Together, these sections provide the methodological and data foundation for the proposed target identification framework. Efficiency modeling presented later in the paper.

2.1 | The Slacks-Based Measure Model

The SBM model represents one of the most advanced developments within the framework of DEA, designed to overcome the limitations of classical models such as CCR and BCC [7]. Traditional DEA models operate on a radial approach, assessing inefficiency through proportional reductions in inputs or expansions in outputs. However, in practice, many DMUs exhibit non-radial inefficiencies, which manifest as slacks-excesses in inputs or shortfalls in outputs. The SBM model explicitly incorporates these slacks into its efficiency measurement, thereby providing a more accurate and comprehensive assessment.

A key feature of the SBM model is its direct inclusion of input and output slacks in the objective function. This enables the model to account for all deviations from the efficient frontier. In contrast, classical DEA models may erroneously classify a unit as efficient if its radial efficiency score is close to one, even when significant slacks exist in some inputs or outputs.

The SBM model rectifies this by integrating all forms of inefficiency into the final efficiency score, thus preventing the overestimation of performance. From a mathematical standpoint, the objective function of SBM is constructed to minimize a fraction composed of the normalized sum of input and output slacks. The model can be defined in input-oriented, output-oriented, or non-oriented forms, offering flexibility in its application.

Furthermore, SBM is not restricted to relative comparisons; it can also uncover behavioral patterns in data, particularly in environments where performance is affected by structural or operational constraints.

The application of the SBM model is particularly prominent in fields such as healthcare, finance, education, and industry. For example, in analyzing patient data, SBM can evaluate which patients, given specific biological characteristics, achieve better treatment outcomes. This capability allows healthcare professionals and decision-makers to optimize treatment paths and allocate medical resources more efficiently.

In addition, SBM provides efficiency scores less than or equal to one, facilitating a more precise ranking of DMUs. Unlike some DEA models that simply differentiate between efficient and inefficient units, SBM quantifies the degree of inefficiency. This nuanced assessment enables more targeted goal-setting and the development of effective improvement strategies for underperforming units.

In hybrid analytical approaches, SBM is frequently used to generate efficiency labels that serve as input for machine learning models such as decision trees, neural networks, or clustering algorithms. This integration of DEA and machine learning enables the discovery of hidden patterns within complex and multidimensional datasets, opening new avenues in data-driven decision support systems.

Another important aspect of SBM is its robustness in handling noisy data and variables with differing scales, due to its non-radial and non-oriented nature. Empirical studies have demonstrated that SBM is less sensitive to outliers compared to other DEA models, offering greater reliability in real-world data analysis. This robustness has made SBM a preferred model among researchers for empirical efficiency studies.

In summary, the SBM model is not only a tool for accurately assessing the relative efficiency of units, but also a foundation for deeper analysis of inefficiencies, the identification of optimal performance patterns, and the design of improvement policies. These features have established SBM as a core element in the development of efficiency and productivity analysis frameworks across diverse scientific disciplines.

The SBM model, proposed by Tone [5], evaluates the efficiency of a DMU by directly incorporating input and output slacks into the efficiency score. The basic SBM model under VRS is formulated as follows:

$$\mathrm{Min}\left(1-(1/\mathrm{m})\frac{\mathrm{T}}{\mathbf{X_o}}\right).\left(1+(1/\mathrm{s})\frac{\mathrm{S}}{\mathbf{Y_o}}\right)^{-1},$$

$$\mathrm{s.t.}$$

$$\sum\nolimits_{\mathrm{j}=1}^{\mathrm{n}}\lambda_{\mathrm{j}}\mathrm{X_j}+\mathrm{T}=\mathrm{X_o},$$

$$\sum\nolimits_{\mathrm{j}=1}^{\mathrm{n}}\lambda_{\mathrm{j}}\mathrm{Y_j}-\mathrm{S}=\mathrm{Y_o},$$

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

$$\lambda_j \geq 0, \text{for all j.}$$

2.2 | The Breast Cancer Dataset

The breast cancer dataset is one of the most well-known and widely used datasets in the fields of machine learning and data science, available through the Scikit-learn library [8]. Originally derived from the Wisconsin Diagnostic Breast Cancer (WDBC) database, this dataset contains clinical and microscopic imaging data from breast tissue samples. The primary aim of the dataset is to distinguish between benign and malignant tumors by analyzing cellular features.

The features extracted from this dataset are based on digital imaging of cell nuclei obtained from biopsy samples. For each case, thirty numerical features are calculated, describing various physical properties of the cell nuclei such as radius, texture, perimeter, area, smoothness, concavity, concave points, symmetry, and fractal dimension.

These features are derived from different statistical aspects namely the mean, standard error, and "worst" (Largest) values-offering a comprehensive representation of tumor morphology. One of the strengths of this dataset is the relatively balanced proportion of benign and malignant cases, making it especially suitable for binary classification tasks.

The dataset contains 569 instances with 30 standardized numerical features, and each sample is labeled as either benign or malignant. This simple yet powerful structure makes the dataset ideal for developing machine learning models, statistical analysis, and even more advanced techniques such as DEA.

From a medical standpoint, the use of this dataset can contribute to the development of decision-support systems for physicians. Models trained on this data can analyze cellular features and predict malignancy with high accuracy. Such models not only accelerate the diagnostic process but also reduce the risk of misdiagnosis and support more informed clinical decisions.

In the context of DEA, the breast cancer dataset provides a unique opportunity to evaluate the performance efficiency of patients from a biological perspective. Each patient can be treated as a DMU, where the cellular characteristics act as inputs and outputs in DEA models such as the SBM. This enables researchers to identify efficient and inefficient patients and derive insights that could lead to more effective treatment strategies. Due to its numeric structure and absence of missing values, the dataset is highly suitable for various types of statistical and supervised or unsupervised learning analyses.

The features also exhibit significant internal correlations, which, while posing challenges such as multicollinearity, create opportunities for deeper analytical methods like dimensionality reduction, feature selection, and clustering.

The breast cancer dataset has been extensively used in academic research and machine learning competitions such as those on Kaggle, serving as a benchmark for evaluating model performance. High accuracy has been reported with models such as Support Vector Machines (SVMs), random forests, and neural networks. However, combining this dataset with decision-making methodologies such as DEA or AHP enables interdisciplinary analyses that go beyond what machine learning alone can achieve.

Ultimately, the breast cancer dataset from Scikit-learn is not only a classical and educational resource but also a powerful foundation for testing research hypotheses, developing intelligent diagnostic systems, and designing novel analytical approaches at the intersection of data science and medicine. A thorough analysis of this dataset can reveal new insights into cellular patterns, disease progression, and optimization of breast cancer treatment strategies.

2.3 | Problem Statement

In recent years, significant advancements have been made in the field of machine learning and medical data analysis. However, most existing methods primarily focus on classification or prediction, overlooking more structural approaches such as the assessment of individual biological efficiency. Although the breast cancer dataset has been widely applied in classification models, it also offers potential for a novel analytical perspective-provided that a suitable and robust framework is adopted to capture the underlying inefficiencies within the data.

Classical DEA models, such as CCR and BCC, often fail to handle complex biological data effectively due to their simplifying assumptions and inability to account for non-radial inefficiencies. To address these limitations, the SBM was introduced as a more advanced, non-radial model capable of capturing inefficiencies resulting from input excesses or output shortfalls.

Nevertheless, even the standard SBM model may fall short when applied to high-variability medical datasets, such as those in oncology. This has led to the development of modified SBM models, which offer greater flexibility and precision in modeling real-world inefficiencies.

Given the numerical structure and correlated features of the breast cancer dataset, the use of a modified SBM model provides an innovative solution for evaluating the biological efficiency of each patient-beyond the conventional binary classification into benign or malignant cases.

This approach enables the identification of efficient and inefficient patients and helps uncover influential patterns among cellular features. As a result, the analysis shifts from mere prediction to biological benchmarking, offering new avenues for clinical insight and optimized decision-making.

2.4 | Research Contribution

This study employs a modified SBM model to evaluate the biological performance of patients in the Scikit-learn breast cancer dataset and presents contributions across three main dimensions:

I. First, it introduces a novel framework by modeling patients as DMUs, using cellular features as inputs and outputs within the DEA structure-bridging medical data with efficiency analysis.

II. Second, the use of a modified SBM enables the direct and accurate detection of non-radial inefficiencies that are often overlooked in traditional DEA models.

III. Third, by analyzing the efficiency scores, the study identifies reference patients and benchmarks, allowing for biological pattern recognition and the suggestion of optimal treatment pathways.

This research creates an interdisciplinary link between data science, operations research, and medicine, and demonstrates the untapped potential of DEA in clinical data analysis.

3 | Methodology

To effectively evaluate patient efficiency in breast cancer diagnostics, this study integrates machine learning with operational research techniques. Initially, a Decision Tree algorithm is employed to identify the most significant clinical features. These findings then inform the development of a modified SBM model under the VRS assumption, enabling more accurate and interpretable efficiency analysis.

3.1 | Feature Selection Using Decision Tree on Breast Cancer Data

Understanding which features are most influential in distinguishing between benign and malignant tumors is crucial in any predictive or analytical study in oncology. Machine learning algorithms offer powerful tools to uncover such patterns in data. Among them, Decision Tree algorithms stand out due to their interpretability and ability to rank features based on their discriminative power [9], [10].

In this study, we utilize the breast cancer dataset provided by Scikit-learn, which includes 30 numerical features derived from digitized images of breast mass cell nuclei. Each observation corresponds to a patient, and the output label indicates whether the tumor is benign or malignant. These features represent metrics such as radius, texture, smoothness, compactness, and symmetry of the cell nuclei.

We begin by applying a standard Decision Tree classifier to this dataset. The tree is trained using the full set of features and the binary class label as the target variable. During training, the algorithm identifies the features that offer the highest information gain i.e., those that most effectively split the data into homogenous subgroups with respect to the target label.

The trained decision tree is then analyzed to extract the most important features. These are typically found near the root of the tree, where decisions have the most impact on overall classification. Features such as "worst radius," "mean concavity," and "worst perimeter" often emerge as highly significant in classifying the data correctly.

One major advantage of the Decision Tree algorithm is that it inherently handles non-linear relationships and interactions among features. This is especially important in medical data, where the relevance of a feature might not be linear or independent of others.

Hence, this technique not only ranks features but also captures complex interdependencies. After identifying the key features, we normalize and rank them to assign importance weights. These weights will be used later to modify the traditional SBM model, ensuring that more informative variables receive higher consideration during efficiency evaluation.

This process of feature selection serves as a data-driven basis for prioritization in the efficiency model. Rather than relying on expert-driven or subjective feature importance, the Decision Tree offers an objective measure based on actual performance in classification. The interpretability of decision trees also allows clinicians or healthcare analysts to validate and understand which variables drive the classification process, fostering trust in the analysis. This transparency is often lacking in more complex or black-box models such as neural networks.

Overall, this initial stage not only reduces dimensionality and computational complexity for the DEA model but also strengthens its clinical relevance by grounding the analysis in medically significant variables.

This integrated feature selection step represents a critical methodological bridge between machine learning and operational research, paving the way for the development of a robust, targeted, and interpretable efficiency evaluation model.

3.2 | Modified Slacks-Based Measure Model under Variable Returns to Scale

The SBM model is a non-radial, non-oriented DEA model that evaluates the efficiency of DMUs by explicitly accounting for input excesses (Slacks) and output shortfalls. It was originally introduced by Tone [5] and has since become a valuable tool in performance measurement, particularly in healthcare settings.

In traditional DEA models, all features are treated equally unless otherwise specified. However, in domains such as medical diagnostics, not all variables have the same clinical relevance. Incorporating this variation in importance is crucial to avoid biased or misleading efficiency results.

To address this, we propose a modified SBM model that integrates feature importance scores derived from the Decision Tree analysis into the efficiency evaluation. The key idea is to assign weights to each input feature proportional to its predictive power in classifying tumor type. This allows the model to focus more on clinically significant variables. Our model operates under the VRS assumption.

This assumption is more realistic in healthcare and medical contexts where scale efficiency is not constant across all units (i.e., patients). Under VRS, the production possibility set is convex but not necessarily proportionate, allowing for more flexibility in benchmarking.

Mathematically, we adapt the original SBM objective function by applying feature-specific weights in the numerator (Input slacks) and optionally in the denominator (Output slacks if outputs are included). This results in a weighted slack-based efficiency score that better reflects the relative importance of each variable. Each patient is modeled as a DMU, with the selected features as inputs and an optional output such as diagnostic accuracy, risk index, or survival proxy. The model computes the efficiency of each DMU by measuring its distance from the efficient frontier formed by the best-performing patients, excluding itself.

The efficiency score ranges from 0 to 1, with values closer to 1 indicating higher efficiency. Inefficient patients are those with excessive inputs (i.e., poor clinical indicators) relative to the best-performing peers with similar characteristics. This modified SBM model is particularly useful for identifying target patients-those who are consistently efficient and can serve as references for treatment planning or further investigation.

Conversely, inefficient patients may be candidates for special care or clinical review. By combining machine learning with DEA, we create a hybrid model that benefits from both data-driven feature discovery and mathematically rigorous efficiency evaluation. This enhances not only the accuracy but also the practical interpretability of the results. The integration of VRS in the model further improves its adaptability to real-world clinical data, where patient heterogeneity and differences in biological profiles are common and must be acknowledged.

In summary, the modified SBM model under VRS, guided by Decision Tree-based feature weighting, provides a comprehensive and meaningful framework for assessing efficiency and discovering target patterns in breast cancer data. In the proposed modified SBM model, the weights assigned to inputs (v) and outputs (u) are not treated as equal or arbitrary. Instead, they are determined based on the feature importance scores obtained from a decision tree algorithm.

The decision tree is first applied to the breast cancer dataset to identify the most influential features in distinguishing between benign and malignant tumors. These relative importance values are then used to assign weights in the objective function of the SBM model. In this way, statistical insights and classification structure from the decision tree are directly integrated into the DEA framework.

The modified SBM model, using these importance-based weights, enables a more accurate and targeted evaluation of DMUs. The objective is to minimize the weighted sum of input and output slacks, with the modification that each variable's contribution is scaled according to its significance in cancer classification as determined by the decision tree.

This approach enhances the scientific validity of the model and improves the quality of pattern recognition, since inputs and outputs with greater diagnostic relevance receive higher emphasis in the efficiency assessment. The complete formulation of the modified model is presented in the

following.

$$\mathrm{Min}\left(1-(1/\mathrm{m})\frac{\mathrm{VT}}{\mathrm{X_o}}\right).\left(1+(1/\mathrm{s})\frac{\mathrm{US}}{\mathrm{Y_o}}\right)^{-1},$$

$$\mathrm{s.t.}$$

$$\sum_{\mathrm{j}=1}^{\mathrm{n}}\lambda_{\mathrm{j}}\mathrm{X_j}+\mathrm{T}=\mathrm{X_o},$$

$$\sum_{\mathrm{j}=1}^{\mathrm{n}}\lambda_{\mathrm{j}}\mathrm{Y_j}-\mathrm{S}=\mathrm{Y_o},$$

$$\sum_{\mathrm{j}=1}^{\mathrm{n}}\lambda_{\mathrm{j}}=1,$$

$$\lambda_{\mathrm{j}}\geq 0,\text{for all j.}$$

In the above model, V and U are vectors that represent the priorities in the objective function, particularly with respect to the input and output slack variables.

4 | The Breast Cancer Wisconsin Dataset

The breast cancer Wisconsin dataset comprises information on 569 samples from female patients, with each sample described by 30 numerical features. These features represent geometric and statistical characteristics of cell nuclei obtained through digital imaging techniques of fine needle aspirates. Additionally, each sample is labeled to indicate whether the tumor is benign or malignant.

An initial analysis shows that out of the 569 available samples, 357 belong to the benign class and 212 to the malignant class. This proportion indicates a relatively balanced dataset, though the imbalance may still influence the performance of classification models. Statistically, this distribution suggests that malignant cases account for approximately 37% of the samples.

The overall mean of the "mean radius" feature across the dataset is approximately 14.13, with a range spanning from about 6 to over 28. This indicates a wide variation in tumor size, suggesting that size is potentially a key factor in determining malignancy. The standard deviation of this feature is 3.52, reflecting a moderate level of dispersion within the data.

The "mean area" feature has the highest average value among all features, with a mean of approximately 654.89. Its standard deviation is also quite high (351.9), indicating significant variability in cell area among the samples. This feature is likely to be a strong discriminator between benign and malignant tumors.

Among the features related to texture, the “mean texture” has an average of 19.3 and a standard deviation of 4.30. Compared to other features, this is relatively low, possibly indicating a lesser role in classification; however, further analysis using machine learning models is needed to confirm this assumption.

When data is analyzed by class, it is observed that the average radius of malignant tumors is significantly larger than that of benign ones. Specifically, the average radius for malignant samples is approximately 17.5, whereas for benign samples it is around 12.1. This statistically significant difference highlights the importance of this feature in cancer classification models.

Other features, such as “compactness” and “concavity,” also show higher mean values in malignant tumors than in benign ones. For example, the average compactness in malignant tumors is around 0.16, compared to 0.08 in benign tumors. These patterns suggest that malignant tumors are not only larger but also have more complex geometric structures.

Conversely, some features such as “symmetry” and “fractal dimension” show little difference between the two classes. For instance, the average symmetry values are quite similar for both benign and malignant samples. Such features might have limited impact in classification and are often excluded in feature selection processes.

Furthermore, the statistical analysis indicates that certain features are highly correlated, such as mean radius, mean area, and mean perimeter. These correlations can be useful in identifying clusters of related features and in reducing redundancy. Techniques like dimensionality reduction or selecting uncorrelated features can improve model efficiency.

Overall, the initial statistical analysis of the breast cancer dataset provides a strong foundation for further machine learning applications and efficiency assessment models like the SBM. In subsequent stages, this information can guide the selection of relevant features and enhance the design of predictive models and efficient target identification.

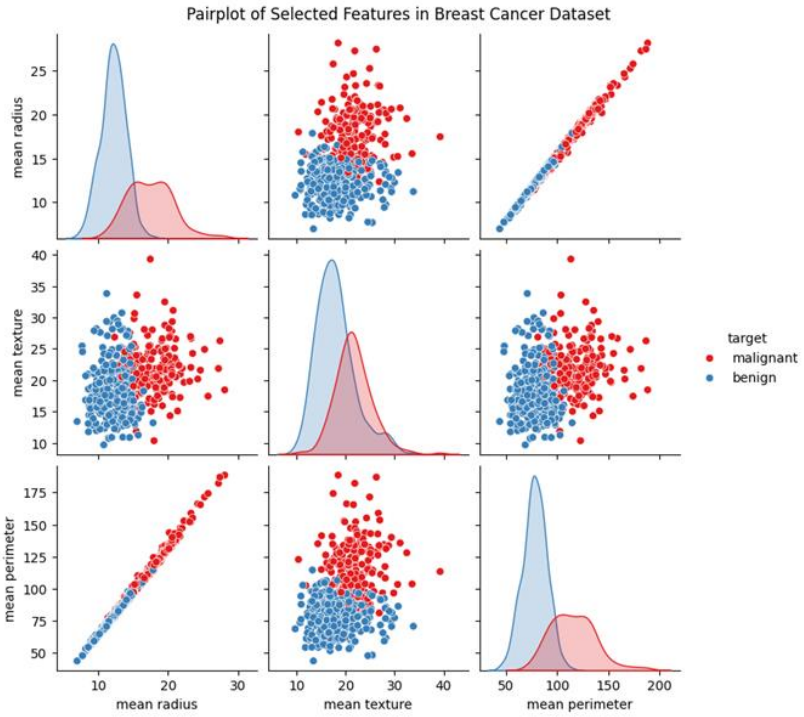

Fig. 1. illustrates the pairwise scatterplots and density distributions for three key features in the breast cancer dataset.

Fig. 1 illustrates the pairwise scatterplots and density distributions for three key features in the breast cancer dataset: mean radius, mean texture, and mean perimeter. Each point in the scatterplots represents a single patient record, color-coded by tumor type-red for malignant and blue for benign. The diagonal of the matrix displays the Kernel Density Estimates (KDE) of each individual feature, separated by class.

In the scatterplot between mean radius and mean perimeter, a strong positive correlation is observed, particularly among malignant cases. Malignant tumors tend to exhibit larger values for both features, clustering in the upper-right quadrant of the plot. Conversely, benign tumors are concentrated in the lower-left region, indicating that smaller radius and perimeter values are typically associated with non-cancerous cases. This separation suggests that these two features are highly discriminative and effective for classifying tumor malignancy.

The plot comparing mean radius and mean texture also displays class-specific patterns. While the distinction is not as sharp as in the previous plot, there is a general trend showing malignant tumors with higher radius and texture values. This confirms the complementary role of texture in supporting classification, especially when used alongside size-related features such as radius.

The scatterplot between mean perimeter and mean texture reveals a similar pattern. Although some overlap exists between the two classes, malignant tumors generally occupy areas with larger perimeter and moderate-to-high texture values. Benign tumors again show lower values across both dimensions. This indicates that perimeter and texture, while individually less decisive than radius, contribute valuable information when analyzed together.

The KDE plots along the diagonal further reinforces these findings. In each case, malignant and benign tumors show clearly distinct density peaks. Malignant cases tend to skew toward higher values, while benign cases cluster around lower ranges. These density plots support the hypothesis that the selected features are statistically separable and thus suitable for downstream analysis using decision trees and DEA models.

Overall, *Fig. 1* provides a strong visual confirmation of the discriminative power of the selected features. Their integration into classification and efficiency assessment models such as a modified SBM is therefore justified both statistically and clinically.

The numerical efficiency scores obtained from the modified SBM model reveal significant insights into the relative performance of the DMUs under evaluation. Out of 446 DMUs, 98 units (Approximately 21.97%) achieved a full efficiency score of 1.000, indicating that these units are operating on the efficiency frontier without any input excesses or output shortfalls.

These efficient units can be considered as benchmarks or role models for the remaining DMUs. The remaining 348 DMUs (About 78.03%) are inefficient to varying degrees. The lowest recorded efficiency score is 0.6351, and a notable number of DMUs have scores falling between 0.65 and 0.85, indicating considerable room for improvement in resource utilization or output generation. This distribution suggests a relatively wide performance gap across the dataset.

The majority of DMUs fall in the moderate efficiency range (0.70–0.90), highlighting systemic inefficiencies which might stem from operational practices, resource allocation, or external constraints. It is also important to note that even among the inefficient units, many exhibit scores close to the efficiency frontier (e.g., above 0.90), suggesting that minor improvements could render them efficient.

The findings underscore the practical utility of the modified SBM model in not only distinguishing efficient and inefficient units but also in identifying specific improvement potentials. Given that the model accounts for both input excesses and output shortfalls, the results reflect a nuanced and realistic measure of efficiency that is highly relevant for performance optimization and strategic planning in complex operational settings.

5 | Conclusion

This study introduces a novel integrative approach that combines a modified SBM with a Decision Tree algorithm to enhance the analysis of breast cancer data. By treating individual patients as DMUs, we were able to assess their relative efficiency based on critical clinical features, while also addressing the limitations of traditional DEA models that often fail to differentiate between equally efficient units.

The Decision Tree algorithm played a crucial role in identifying the most discriminative features-such as mean radius, mean texture, and mean perimeter-which were subsequently emphasized in the SBM model to improve the sensitivity of efficiency analysis. Visual inspection through scatterplots and KDE distributions confirmed the strong correlation between these features and tumor type, thus validating their clinical relevance and statistical separability.

The Modified SBM model, equipped with these key feature weights, successfully categorized patients into efficient and inefficient groups. Approximately 22% of the DMUs were found to be fully efficient, serving as performance benchmarks, while the remaining 78% displayed varying degrees of inefficiency. This stratification offers valuable insights into performance disparities and highlights potential areas for clinical or operational improvement.

Importantly, the combined method offers not just classification but also interpretation-enabling the identification of biologically significant "target" patients who exhibit distinct efficiency profiles. These profiles can aid in personalized treatment planning, resource allocation, and early intervention strategies. In conclusion, the integration of modified SBM and Decision Tree algorithms provides a powerful, data-driven framework for uncovering latent patterns in medical datasets. This hybrid model holds promise for broader applications in healthcare analytics, particularly in enhancing diagnostic precision and supporting evidence-based decision-making in cancer care.

Conflict of Interest

The authors declare no competing interests.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Funding

No funding was received for conducting this study.

References

[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] Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European journal of operational research*, *130*(3), 498–509. https://doi.org/10.1016/S0377-2217(99)00407-5

[4] 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

[5] Tone, K. (2002). A strange case of the cost and allocative efficiencies in DEA. *Journal of the operational research society*, *53*(11), 1225–1231. https://doi.org/10.1057/palgrave.jors.2601438

[6] Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management science*, *39*(10), 1261–1264. https://doi.org/10.1287/mnsc.39.10.1261

[7] Lee, H. S. (2021). An integrated model for SBM and super-SBM DEA models. *Journal of the operational research society*, *72*(5), 1174–1182. https://doi.org/10.1080/01605682.2020.1755900

[8] Mohi ud din, N., Dar, R. A., Rasool, M., Assad, A. (2022). Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. *Computers in biology and medicine*, *149*, 106073. https://doi.org/10.1016/j.compbiomed.2022.106073

[9] Lavanya, D., & Rani, D. K. U. (2011). Analysis of feature selection with classification: Breast cancer datasets. *Indian journal of computer science and engineering (IJCSE)*, *2*(5), 756–763. https://ijcse.com/docs/INDJCSE11-02-05-167.pdf

[10] Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. *2011 IEEE control and system graduate research colloquium* (pp. 37–42). IEEE. https://doi.org/10.1109/ICSGRC.2011.5991826