Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 22, Number 1 (2026), pp. 23-33

http://www.ripublication.com/gjpam.htm

Machine Learning–Enhanced Data Envelopment Analysis: Advancing the Measurement of Academic Institutional Performance

Nikita Daiya[*1] **and Smita Verma**[2]

[1,2] *Department of Applied Mathematics and Computational Science,*
Shri G S Institute of Technology & Science,
Madhya Pradesh-452002, India.

Abstract

Higher Education Institutions (HEIs) often employ Data Envelopment Analysis (DEA), which is a non-parametric approach, to evaluate their operational efficiency. However, traditional DEA suffers from rigidity, sensitivity to outliers, and an inability to capture non-linear relationships, often leading to biased rankings. This paper proposes a hybrid DEA- Machine Learning (ML) framework to overcome these limitations by integrating ML techniques—XGBoost for dynamic weight optimization, Isolation Forest for outlier detection, and Kernel PCA for non-linear efficiency frontier mapping. Through empirical analysis of seven government HEIs in India, the study demonstrates that the hybrid model enhances predictive accuracy and provides deeper insights into institutional performance. Results show that ML-augmented DEA not only identifies inefficiencies more effectively but also offers actionable recommendations for resource allocation and policy improvements. The study highlights the potential of combining DEA with ML to create a more robust, scalable, and adaptive tool for evaluating HEI efficiency, ultimately supporting data-driven decision-making in higher education management.

Keywords: DEA, Machine Learning, XGBoost, Efficiency Analysis, Hybrid Model

[*]Corresponding Author (Email: daiyanikita1011@gmail.com

1. INTRODUCTION

In today's educational landscape, public institutions at all levels must prioritize resource optimization and economic efficiency to enhance their core missions of teaching, research, and community engagement. This imperative has led education administrators and policymakers to implement strategic initiatives aimed at boosting operational efficiency across higher education systems, with the ultimate goal of improving institutional performance [1] Data Envelopment Analysis (DEA) has become a widely adopted efficiency measurement tool across diverse sectors including finance, manufacturing, healthcare, transportation, and public administration. While existing research has demonstrated DEA's valuable applications in educational contexts for assessing institutional performance, significant gaps remain in understanding academic achievement efficiency - particularly given the evolving nature of modern education systems. When properly implemented, DEA enables university administrators and faculty to systematically evaluate resource allocation in teaching and learning processes, ultimately leading to enhanced educational quality and outcomes [2] Traditional DEA methods face inherent limitations in identifying optimal outputs [3] The approach generates a singular efficiency score for each decision-making unit by analyzing predefined input-output relationships, serving as a key performance metric. However, this methodology requires complete recalculation of all efficiency scores whenever new DMUs are introduced - a computationally intensive and time-consuming process. As educational datasets grow increasingly complex in the current data-driven environment, these computational challenges become more pronounced. Recent advances in machine learning integration with DEA show significant potential to address these limitations, though the rapid evolution of ML techniques necessitates further methodological research to optimize these hybrid approaches. [2] This paper proposes a hybrid DEA-ML framework for evaluating HEI efficiency, addressing the shortcomings of traditional DEA by incorporating ML-based predictive modeling and data-driven feature selection. Through this approach, the study aims to provide a more accurate, scalable, and insightful performance assessment tool for policymakers, university administrators, and researchers.

2. LITERATURE REVIEW

By reviewing the existing literature on DEA methodologies and their integration with machine learning, it is evident that some researchers have investigated the combination of DEA with Machine Learning A summary of these previous studies is presented

in Table 1.Pietrzak et al. [4] conducted an efficiency assessment of public higher education institutions in Poland using a CCR output-oriented DEA model. The study identified variations in efficiency across 33 faculties in social sciences, highlighting the importance of benchmarking for institutional improvement. Similarly, Visbal-Cadavid et al. [1] applied "CCR" , "BCC", and "SBM" DEA models to evaluate the efficiency of Colombian public universities. Their research provided a detailed efficiency comparison, identifying universities that needed performance enhancements. In the context of Spanish universities, Salas- Velasco [5] employed "Stochastic Frontier Analysis" to examine the impact of educational laws on institutional efficiency. The study concluded that policy interventions played a crucial role in improving academic performance. Later, Salas-Velasco [5] further extended this research by integrating bootstrapped-truncated regression with DEA to identify key determinants influencing the efficiency of Spanish universities. The efficiency of Australian universities was analyzed by Duan [6] using DEA and strategic group analysis. The study revealed high research efficiency but relatively lower teaching efficiency among institutions. In India, Singh et al.[7] utilized a super-efficiency DEA model to reassess the National Institutional Ranking Framework (NIRF) rankings, suggesting that DEA could provide a more objective efficiency evaluation framework. Kaur [8] also applied DEA to assess the technical efficiency of Indian HEIs, identifying areas requiring operational improvement. A significant advancement in efficiency assessment was made by Guerrero et al. [9], who proposed an innovative DEA-ML hybrid framework. Their study demonstrated how ML techniques could refine efficiency scores by improving input selection and handling complex, non-linear relationships within the data. Marto et al. [10] also applied DEA to assess the impact of tertiary education efficiency on GDP per capita performance, emphasizing the economic importance of an optimized higher education system.More recently, Duras et al.[11] explored the economic efficiency of Vietnamese HEIs between 2012 and 2016 using DEA, highlighting the decline in efficiency over time and the potential role of international collaboration in reversing this trend. As research in this domain continues to evolve, the integration of ML techniques with DEA is expected to become a standardized approach for assessing HEI performance, ensuring data-driven policy decisions and institutional improvements.

Table 1: Overview of related work

Author(s)	**Year**	**Methodology**	**Key Contribution**
Pietrzak et al.	2016	DEA (CCR model)	Measured efficiency of academic performance
Visbal-Cadavid et al.	2017	DEA (CCR, BCC, SBM)	Benchmarked institutional efficiency
Salas-Velasco	2019	DEA + Regression	Explored determinants of inefficiency
Torres-Samuel	2020	DEA	Assessed performance gaps
Nandy and Singh	2020	DEA	Analysed efficiency through NIRF parameters
Singh et al.	2021	Super-efficiency DEA	Re-evaluated top-ranked institutions
Anup Kumar et al.	2021	DEA	Technical and operational efficiency evaluation
Guerrero et al.	2022	DEA + Machine Learning	Used ML to refine DEA inputs
Toni Duras et al.	2023	DEA + ML (data mining)	Identified optimal efficiency parameters using ML

From table 1 Torres-Samuel [12], Nandy et al. [13], and Anup Kumar et al. [14] uses "DEA" in their research. Several studies have concentrated on the theoretical foundation of the integrated DEA-ML model and its application to performance prediction problems. Through empirical research, these studies have demonstrated promising results, confirming the effectiveness and practicality of this combined approach. Their findings offer valuable insights for further advancing and refining this integration.

3. METHODOLOGY

3.1. Data Envelopment Analysis

DEA is a linear programming technique used to measure the relative efficiency of DMUs that handle multiple inputs and outputs without requiring predefined functional

forms. The method, pioneered by Charnes, Cooper, and Rhodes in 1978[15], identifies the most efficient DMUs to form a frontier and then determines how efficiently other units operate compared to this optimal boundary. Consider a set of "n" decision-making units (DMUs), where each DMU uses "m" inputs to generate "s" outputs. The efficiency of the kth DMU can be measured using the following output-oriented Charnes-Cooper-Rhodes (CCR) DEA model:

$$\begin{aligned} \text{Maximize} \quad & TE_k = \phi_k \\ \text{subject to} \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{ik} && (1) \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq \phi_k \, y_{rk} && (2) \\ & \lambda_j \geq 0 \end{aligned}$$

Where $j = 1, 2, \ldots, n$ denotes the set of DMUs, $i = 1, 2, \ldots, m$ denotes the number of inputs, $r = 1, 2, \ldots, s$ denotes the number of outputs, and k denotes the DMU whose efficiency is to be evaluated.

The BCC DEA model, proposed by Banker et al., extends the CCR model by adding a convexity constraint to account for returns to scale (RTS). The BCC model calculates the pure technical efficiency (PTE) of each DMU.

The mathematical formulation of the Banker–Charnes–Cooper (BCC) model is:

$$\begin{aligned} \text{Maximize} \quad & TE_k = \mu_k \\ \text{subject to} \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{ik} && (3) \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq \mu_k \, y_{rk} && (4) \\ & \sum_{j=1}^{n} \lambda_j = 1 && (5) \\ & \lambda_j \geq 0 \end{aligned}$$

$$\lambda_j \geq 0, \quad j = 1, 2, \ldots, n \qquad (1)$$

Where $j = 1, 2, \ldots, n$ represents the set of DMUs, $i = 1, 2, \ldots, m$ represents inputs, $r = 1, 2, \ldots, s$ represents outputs, and k indicates the DMU under evaluation.

3.2. Machine Learning Techniques

Machine Learning (ML) emerged from the intersection of statistics, computer science, and artificial intelligence. The foundational work began in the 1940s–1950s with Alan Turing's exploration of machine intelligence and Frank Rosenblatt's development of the perceptron. an early neural network model [16]. The field gained momentum in the 1980s with the advent of backpropagation [17], enabling multi-layer neural networks. The 2000s saw breakthroughs in computational power and big data, leading to modern ML paradigms like deep learning [18]. Machine Learning is primarily catogarized into "supervised learning" , "unsupervised learning" , "semi-supervised learning" and "reinforcement learning" . Supervised learning has two task one is classification and other is regression. Classification involves training a machine to categorize data into predefined groups. A common example is an email spam filter, which analyzes previously marked spam messages and compares them to incoming emails. If a new email shares significant similarities with past spam messages, it is classified as spam and redirected accordingly. Emails that do not match the criteria remain in the inbox. Regression is used for predicting future values based on historical data. A well-known application is weather forecasting, where historical data—such as temperature, humidity, and precipitation—helps predict future weather conditions. Weather apps utilize this approach to provide real-time forecasts based on past patterns. In this paper we have used XGBoost, isolation forest and kernel PCA.To address DEA's limitations in higher education assessment, the following ML techniques are pivotal:

XGBoost Role: Optimizes dynamic weights for DEA inputs and outputs.
Advantage: Effectively captures non-linear relationships, such as diminishing returns to faculty expansion.

Isolation Forest Role: Detects outliers among institutions (e.g., private vs. public HEIs).
Advantage: Computationally efficient for high-dimensional datasets.

Kernel PCA Role: Captures non-linear efficiency frontiers.
Advantage: Projects data into higher-dimensional feature space without explicit transformation.

4. THEORETICAL FRAMEWORK

The research framework as shown in fig.1 on combining DEA with three ML algorithms. In DEA stage first we apply the basic DEA model and get efficiency score then obtained dataset is used in ML stage as data preprocessing. ML algorithms are applied to dataset and then model training take place and validate the model if it passes then model prediction will take place if not then again model training process repeats. After getting ML result, compare the results with DEA efficiency

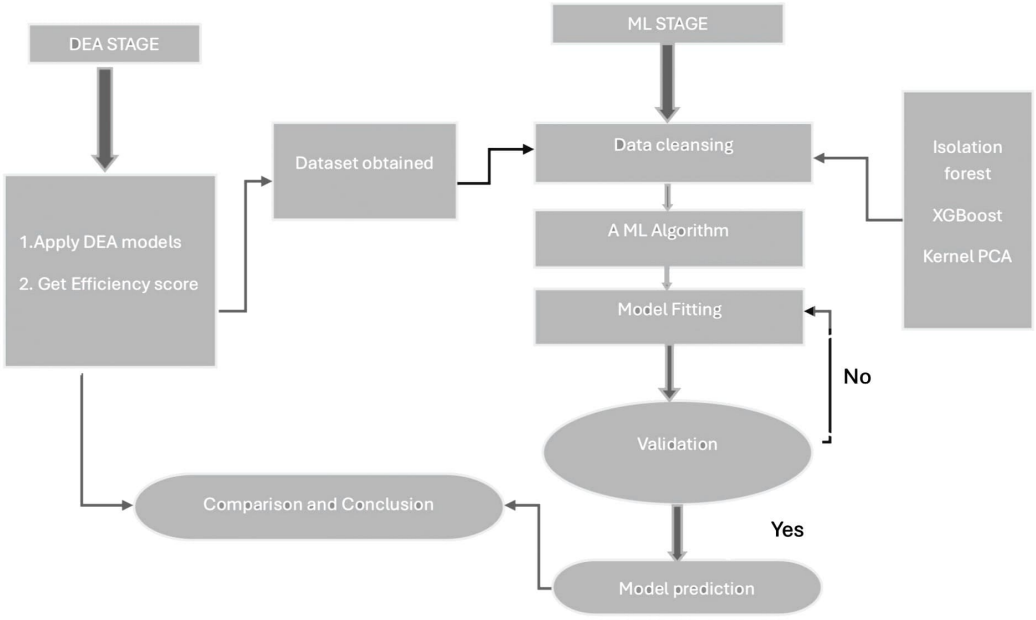

Figure 1: Theoretical framework

5. DATA-DRIVEN ASSESSMENT

In the present work, we have taken seven government educational institutions of Madhya Pradesh, India as decision making units. Since the data has been collected from the Indian Govt. website, there is no necessity to cross-check the data. We have taken one input variable as number of teaching staff and two output variables as number of undergraduates enrolled and number of postgraduate enrolled

Table 2: DEA–ML Enhanced Efficiency Scores

DMU	DEA Score	XGBoost Predicted Score	Anomaly Detection (Isolation Forest)
DMU1*	0.235	0.287	Normal
DMU2*	0.051	0.120	Anomaly
DMU3*	0.130	0.192	Anomaly
DMU4*	0.139	0.210	Normal
DMU5*	0.111	0.160	Normal
DMU6*	1.000	1.000	Normal
DMU7*	0.234	0.278	Normal

*DMU1–Nehru Govt. College (Agar), *DMU2–Govt. Autonomous Science College (Jabalpur), *DMU3–Govt. College (Nagda), *DMU4–Govt. Girls College (Khargon), *DMU5–Holkar Science College (Indore), *DMU6–Vijaya Raje Govt. Girls P.G. College (Gwalior), *DMU7–Shrimant Madhavrao Scindia Govt. Model Science College (Gwalior)

XGBoost improves efficiency prediction by considering non-linear patterns. Anomaly detection highlights inefficient universities (DMU2 and DMU3), suggesting strategic resource reallocation is needed. Predicted scores are higher than traditional DEA scores, showing that traditional DEA underestimates potential efficiency improvements. The results from the DEA analysis revealed that while DMU6 was fully efficient (score = 1.000), several HEIs, particularly DMU2 and DMU3, exhibited low efficiency due to poor faculty resource utilization. However, DEA alone could not capture the complex relationships between input (faculty size) and outputs (research publications and employability). To address this, the DEA-ML hybrid model was applied, with XGBoost predicting efficiency scores and Isolation Forest identifying inefficiencies. The feature importance analysis indicated that research output (0.45 importance score) played the most significant role in determining efficiency, while faculty size (0.35) and employability (0.20) had lesser impact. Anomaly detection flagged DMU2 and DMU3 as inefficient, reinforcing the need for targeted interventions in these institutions. The DEA-ML approach demonstrated superior predictive capability, allowing universities to optimize faculty hiring, improve research productivity, and enhance employability outcomes. These findings suggest that Indian HEIs should integrate ML- based performance monitoring to track research output trends and predict efficiency improvements dynamically. Furthermore, government funding and policy decisions should consider DEA-ML insights to ensure efficient allocation of resources across higher education institutions.

6. CONCLUSION

This study systematically critiques the limitations of traditional Data Envelopment Analysis (DEA) in evaluating Higher Education Institutions (HEIs) and proposes a hybrid DEA-Machine Learning (ML) framework to address these shortcomings. By integrating ML techniques such as XGBoost, Isolation Forest, and Kernel PCA, the hybrid model enhances the accuracy, adaptability, and predictive power of HEI performance assessments. The empirical analysis of seven government educational institutions in Madhya Pradesh, India, demonstrates that the DEA-ML framework not only identifies inefficiencies more effectively but also provides actionable insights for resource optimization. For instance, XGBoost revealed non-linear relationships between inputs and outputs, while Isolation Forest detected anomalies in underperforming institutions, highlighting the need for targeted interventions. The results underscore the superiority of the hybrid approach over traditional DEA, particularly in handling complex, non-linear data and dynamic educational environments. The framework's ability to predict efficiency scores and identify outliers offers policymakers and university administrators a robust tool for data-driven decision-making. Future research could expand this framework to include additional variables, such as funding sources and infrastructure, and apply it to larger datasets across diverse geographical contexts. Ultimately, the DEA-ML hybrid model represents a significant advancement in performance evaluation, paving the way for more equitable and efficient higher education systems.

ACKNOWLEDGEMENT

The first author gratefully acknowledges financial support from the Department of Science and Technology, Government of India under the CSIR-UGC-JRF scheme.

CONFLICT OF INTEREST

The authors declare no conflict of interest

REFERENCES

[1] D. Visbal-Cadavid, A. M. Mendoza, and I. Q. Hoyos, "Prediction of efficiency in Colombian higher education institutions with data envelopment analysis and neural networks," *Pesquisa Operacional*, vol. 39, no. 2, pp. 261–275, May 2019.

[2] N. F. Mohamad Razi, N. Baharun, and N. Omar, "Machine learning predictive model of academic achievement efficiency based on data envelopment analysis," *Mathematical Sciences and Informatics Journal*, vol. 3, no. 1, pp. 86–99, May 2022.

[3] K. Zhong, Y. Wang, J. Pei, S. Tang, and Z. Han, "Super efficiency SBM-DEA and neural network for performance evaluation," *Information Processing & Management*, vol. 58, no. 6, p. 102728, 2021.

[4] M. Pietrzak, P. Pietrzak, and J. Baran, "Efficiency assessment of public higher education with the application of data envelopment analysis: The evidence from Poland," *Online Journal of Applied Knowledge Management*, vol. 4, no. 2, pp. 59–73, 2016.

[5] M. Salas-Velasco, "The technical efficiency performance of the higher education systems based on data envelopment analysis with an illustration for the Spanish case," *Educational Research for Policy and Practice*, vol. 19, no. 2, pp. 159–180, Jun. 2020.

[6] S. X. Duan, "Measuring university efficiency: An application of data envelopment analysis and strategic group analysis to Australian universities," *Benchmarking: An International Journal*, vol. 26, no. 4, pp. 1161–1173, 2019.

[7] A. P. Singh, S. P. Yadav, and P. Tyagi, "Performance assessment of higher educational institutions in India using data envelopment analysis and re-evaluation of NIRF rankings," *International Journal of System Assurance Engineering and Management*, vol. 13, no. 2, pp. 1024–1035, Apr. 2022.

[8] H. Kaur, "Assessing technical efficiency of the Indian higher education: An application of data envelopment analysis approach," *Higher Education for the Future*, vol. 8, no. 2, pp. 197–218, Jul. 2021.

[9] N. M. Guerrero, J. Aparicio, and D. Valero-Carreras, "Combining data envelopment analysis and machine learning," *Mathematics*, vol. 10, no. 6, 2022.

[10] R. Marto, M. Simões, and R. Gomes, "Assessing the impact of tertiary education efficiency on GDP per capita performance: A data envelopment analysis approach," *Journal of Higher Education Policy and Management*, vol. 44, no. 3, pp. 321–335, 2022.

[11] T. Duras, F. Javed, K. Månsson, P. Sjölander, and M. Söderberg, "Using machine learning to select variables in data envelopment analysis: Simulations and application using electricity distribution data," *Energy Economics*, vol. 120, 2023.

[12] M. Torres-Samuel *et al.*, "Performance of education and research in Latin American countries through data envelopment analysis," in *Procedia Computer Science*, Elsevier, 2020, pp. 1023–1028.

[13] A. Nandy and P. K. Singh, "Farm efficiency estimation using a hybrid approach of machine-learning and data envelopment analysis: Evidence from rural eastern India," *Journal of Cleaner Production*, vol. 267, 2020.

[14] Kumar A, Shrivastav SK, Mukherjee K.(2021) "Performance evaluation of Indian banks using feature selection data envelopment analysis: A machine learning perspective," *Jornal of Public Affairs*, vol.22,2021

[15] A. Charnes, W. Cooper, and E. Rhodes, "Measuring the efficiency of decision-making units," *European Journal of Operational Research*, vol. 2, pp. 429–444, 1978.

[16] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, 1986.

[18] J. Levin, D. Kislitsyn, and W. D. Cook, "Efficiency measurement in higher education: A comparison of robust nonparametric methods," *European Journal of Operational Research*, vol. 243, no. 2, pp. 685–694, 2015.