Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 05, No. 02, 2026 ♦ PP. 48–55

Article DOI: https://doi.org/10.54216/IJAIET.050206



FULL LENGTH ARTICLE

A Machine Learning-Enhanced Data Envelopment Analysis Framework for Predicting Institutional Efficiency in Higher Education

Hiteshkumar Solanki[1,*] **Paresh Virparia**[2] **Devika Madalli**[3]

[1] Scientist-D (CS), Information and Library Network (INFLIBNET) Centre, Gandhinagar, India

[2] Professor, P G Department of Computer Science & Technology, Sardar Patel University, Vallabh Vidyanagar, India

[3] Director, Information and Library Network (INFLIBNET) Centre, Gandhinagar, India

Emails: hitesh@inflibnet.ac.in · pvvirparia@yahoo.com · dmadalli@gmail.com

Received: November 07, 2025 Revised: December 17, 2025 Accepted: January 17, 2026 * Corresponding author

ABSTRACT

The paper presents a novel approach for predicting the efficiency of engineering higher educational institutions in India. The prediction model uses data envelopment analysis to solve linear programming problems and support vector regression as a supervised machine learning algorithm. DEA is a nonparametric tool for computing the relative efficiency of HEIs across multiple inputs and outputs. A total of 25 featured variables is considered, out of which 15 are input-oriented, whereas 10 are output-oriented. Input-oriented features are normalised using the standard z-score method, whereas output-oriented features are normalised using the Min–Max normalised method. A sample of 7432 engineering institutions is considered for the featured dataset. The featured dataset is split at a 90:10 ratio for training and testing. The hybrid model, combining traditional DEA with a supervised machine learning SVR, is trained on a training dataset. The prediction model is tested for estimating 743 engineering institutions. The model provides higher accuracy, along with a precision score, in the confusion matrix comparing the actual and predicted HEI performance categories. The best-fitting hybrid model also yields a predicted efficiency index for engineering HEIs, along with MAE, MSE, and RMSE, which are 0.0703, 0.0080, and 0.0894, respectively. The R-squared score of the prediction model is 0.7642. The developed predictive model can be generalised across sectors such as agriculture, banking, industry, and education.

Keywords: Supervised Machine Learning ▪ Data Envelopment Analysis ▪ Relative Efficiency ▪ Support Vector Regression ▪ Prediction Model ▪ Higher Education

1. INTRODUCTION

The Indian Higher Education System is the largest diversified education sector in the world. A total of 1,168 university-level institutions, 45,473 colleges and 12,002 stand-alone institutions are registered in the All-India Survey of Higher Educational Institutions as per the AISHE Report 2021–22 [1]. It offers a wide range of programs, including undergraduate, postgraduate, and doctoral studies, across disciplines such as science, engineering, humanities, management, and vocational education, as per the All-India Survey of Higher Education [1]. Higher educational institutions are classified into several categories, including central universities, state universities, deemed-to-be-universities and private universities, institutes of national importance, affiliated colleges, and stand-alone institutions. Higher education institutions are compromising on infrastructure, manpower, experimental

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

equipment, and additional funding to enhance institutional outcomes. Skill-based and research-oriented courses are being offered under the National Education Policy 2020 [2]. The performance assessment of HEIs is essential for effectively demonstrating their performance and achievements to a wide range of stakeholders, including students, parents, government bodies, funding agencies, regulatory bodies and society. It is based on predefined assessment mechanisms to consider the vital research and academic contributions of faculty members, innovative and impactful research by research scholars, and support staff's efforts to ensure operational engagement throughout the educational system. In a nutshell, this approach offers a relative or absolute, comprehensive representation of institutional growth, contributions, and overall impact on the economic, technical, and societal environment.

The Indian government established ranking and accreditation bodies like the National Assessment and Accreditation Council (NAAC), the National Board of Accreditation (NBA), and the National Institutional Ranking Framework (NIRF) over the past two decades. NAAC was established in 1994 as an independent, accredited council for accrediting HEIs in India. It accredits universities, autonomous colleges, and affiliated colleges through a seven-criterion framework. It uses descriptive, quantitative, and qualitative key indicators-based assessment mechanisms to classify higher educational institutions into seven accredited categories: A++, A+, A, B++, B+, B, and C [3]. The National Institutional Ranking Framework (NIRF) is a critical mechanism established to evaluate and ensure the quality of education services offered by HEIs in India. NIRF ranks institutions annually using five major parameters, including teaching-learning resources, research performance and practices, graduation outcomes, outreach & inclusivity and peer academics & employer perception [4, 5]. Similarly, the NBA focuses on accreditation of professional courses offered by HEIs in the fields of engineering, management, pharmacy, architecture, hospitality & tourism management, computer application, etc.

The study proposes a holistic model for predicting relative efficiency in higher education institutions. This model combines data envelopment analysis with supervised support vector regression to predict the efficiency of each higher education institution. Data Envelopment Analysis is a non-parametric method for assessing Decision-Making Units (DMUs) that accounts for multiple inputs and outputs. The study addresses the limitations of traditional DEA in handling complex, nonlinear relationships among multiple data variables. It also offers a comprehensive tool that supports strategic planning and performance benchmarking for Indian higher education institutions.

2. LITERATURE REVIEW

The DEA-Adaboost model was developed to select efficient suppliers using fuzzy multi-objective DEA in the field of supply chain management. The model was trained using a hybrid approach combining classical DEA with the AdaBoost machine learning algorithm, especially in a big-data scenario [6]. A hybrid model was explored for a combination of data envelopment analysis with eleven machine learning algorithms. A DEA Slack-Based Model (SBM) was conceptualised with an ML algorithm to predict frontier efficiency of carbon emissions [7].

The combined algorithms were developed by combining DEA-CCR with specific algorithms, such as support vector machines, back-propagation neural networks, genetic algorithms with back-propagation neural networks, and improved customised support vector machines. The models were deployed on Chinese manufacturing companies to assess their performance. GANN-DEA was the most accurate model for predicting company efficiency, achieving 94% accuracy [8]. A combined machine learning and data envelopment analysis framework was implemented to predict the continuous efficiency of 500 companies. A dataset of 500 companies was trained with eight machine learning models, including ensemble methods such as CatBoost and LightGBM. The CatBoost model was the most accurate, followed by LightGBM, for the data envelopment analysis with constant returns to scale [9]. A study was conducted to evaluate the performance of various university departments using DEA tools. It was employed across four simulated scenarios involving five departments to enhance the efficiency and quality of services for faculty members' development [10]. An empirical analysis was attempted to integrate relative efficiency analysis techniques, LR, SVM, RF, k-NN, DT and LDA models. An integrated model was implemented across 7548 manufacturing companies in the US, considering Sales as an output data feature. The random forest was considered the best performance model for evaluating companies [11].

A Graph Neural Network (GNN) algorithm with DEA was implemented for predicting annual sales change rates in the field of supply chain management. The model was implemented using flood data and firm-level data, along with various implicit and explicit factors such as annual sales, inter-firm relations, number of employees, and damages incurred in Japan. The study contributed a novel approach that enabled enhanced supply chain management in Industries [12]. A benchmarking analysis was conducted to measure the efficiency of 144 private higher education institutions in Colombia using the Efficiency Analysis Tree (EAT). The EAT and the Convexified frontier counterpart (CEAT) offered the highest discrimination power for identifying efficient HEIs compared with conventional DEA [13]. In the banking sector, a study developed a novel generic framework for classifying DMUs using logistic regression, random forests, and support vector machines. The relative performance index was computed, and each DMU was categorised into groups with labels based on their performance index. The model was trained after applying SMOTE with the Manhattan distance metric to handle an imbalanced banking dataset [14]. A multi-stage model was constructed, with DEA at the initial stage and a machine learning method at the second stage, to identify relevant features. It was used to determine the relative efficiency of 235 European HEIs across 33 features over the 2011–2021 period [15].

Given related studies, developing a prediction model is essential for estimating efficiency for newly added DMUs rather than computing each DMU's relative efficiency using complex, time-consuming procedures.

3. METHODOLOGY

This section outlines the objectives of the proposed study and the feature-variable selection. It also covers the implementation of a research framework for a hybrid model of data envelopment analysis and support vector machine, as described in Figure 1.

3.1 Research Objectives

The foremost objective is to predict the performance category in a category-based assessment of engineering HEIs. The detailed objectives are as follows:

1. To identify the resource-focused and outcome-driven parameters for evaluating engineering institutions.
2. To determine the efficiency benchmark index based on chosen parameters for each engineering institution using DEA.
3. To develop a prediction model for efficiency for new HEI using a machine learning algorithm.

3.2 Architecture of Research Framework

The motive of the proposed study is to develop a combined prediction algorithm by using Data Envelopment Analysis (DEA) and Support Vector Regression (SVR). The proposed research framework consists of three major components, namely 1) Feature Selection and Standardisation, 2) DEA-CCR Model, and 3) SVR-ML Model, as depicted in Figure 1.

3.3 Feature Selection and Standardisation

The preliminary objective of the proposed research is to assess the performance of higher education engineering institutions. To evaluate the performance of higher educational institu-tions that offer engineering courses, the HEIs dataset has been selected, comprising the participating engineering HEIs considered under the India Rankings 2018 to 2024 exercise of the National Institutional Ranking Framework (NIRF).

The NIRF comprises five major parameters: teaching, learn-ing, and resources (TLR); research and professional practices (RP); graduation outcomes (GO); outreach and inclusivity (OI); and perception (PR) [16]. Under each major parameter, the framework is defined in terms of various sub-parameters. Altogether, five major parameters are divided into seventeen sub-parameters for the engineering discipline. A total of 27–30 performance metrics are computed using various absolute or relative formulae to determine weighted scores and ranks of engineering institutions in India [17].

A total of 25 feature variables is considered from the NIRF engineering dataset. All feature variables are grouped into four categories: teaching and learning, financial resources, research outcomes, and graduation outcomes. Out of 25 fea-ture variables, 15 were selected as input features and 10 as output features for the performance evaluation of engineering institutions. The data has been submitted through an online data-capture system within a specific timespan from participating institutions. The previous academic year is considered a timespan for data variables under the teaching-learning cat-egory. Similarly, a three-year time frame has been used for financial resources and for the graduation outcome category. In addition, publications and citations have been retrieved from citation databases such as Clarivate Analytics Web of Science [18] and Elsevier's Scopus [19] for the previous three calendar years for the India Rankings Year. A detailed list of feature variables considered for the performance assessment of higher education engineering institutions is presented in Table 1.

3.4 DEA-CCR Model

A Data Envelopment Analysis (DEA) is a non-parametric tool for measuring the relative efficiency of each Decision-Making Unit (DMU). The DEA addresses multiple input- and output-oriented parameters to solve linear programming problems in operations research. The DEA-CCR model is a frontier analysis model that uses the ratio of weighted multiple outputs to weighted multiple inputs, focusing on each DMU's constraints and ensuring the ratio is no greater than 1. Suppose there is a set of (n) DMUs. Each DMU produces (s) different outputs ((r=1,2,3,...,s)) which are utilising (m) different inputs ((i=1,2,3,...,m)); then the fractional programming model can be expressed as follows:

$$\begin{aligned} \max \quad & \theta = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}, \\ \text{subject to} \quad & \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \le 1, \quad j = 1,2,\ldots,n, \\ & u_r \ge 0, \quad r = 1,2,\ldots,s, \\ & v_i \ge 0, \quad i = 1,2,\ldots,m. \end{aligned} \tag{1}$$

where θ is the efficiency score of the DMU being evaluated; y_{ro} and y_{rj} are output-oriented variables for the evaluated DMU and other DMUs; x_{io} and x_{ij} are input-oriented variables for the evaluated DMU and other DMUs; and u_r and v_i are weights for the output-oriented and input-oriented variables, respectively.

3.5 SVR-RBF Model

The Support Vector Machine (SVM) is a classification algorithm that assigns a class label to a set of categorical variables based on input features. Support Vector Regression (SVR) is a regression technique that extends the concepts of SVMs to predict continuous values rather than class labels for the target variable. The goal of SVR is to generalise a function that approximates the underlying relationship between input features and a continuous target variable with minimal residuals. The SVR-RBF model aims to fit the data within a specified tolerance, using a Radial Basis Function (RBF) kernel to handle non-linear relationships by mapping input features into a higher-dimensional space. The mathematical model of SVR-RBF is expressed as

$$f(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2\right) + b. \tag{2}$$

Here, $f(\mathbf{x})$ is the predicted value of the targeted feature variable; α_i and α_i^* are Lagrange multipliers determined by the optimisation process associated with its support vectors; $\mathbf{x}$ is the input vector for which the prediction is made; $\mathbf{x}_i$ is the (i)-th support vector from the training dataset;

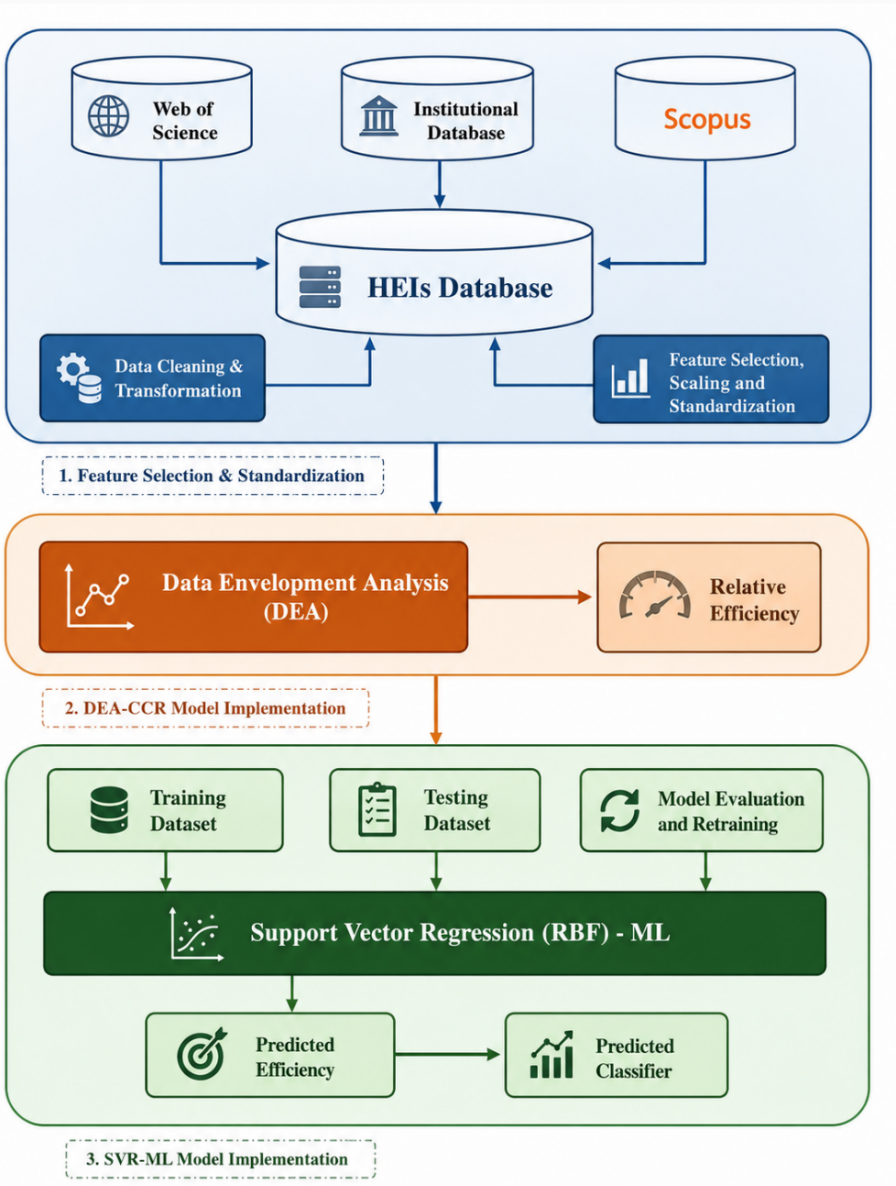


Figure 1. Research framework for predicting efficiency of HEIs using the DEA–SVR model.

$\|\mathbf{x}-\mathbf{x}_i\|^2$ is the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{x}_i$; $\exp(-\gamma\|\mathbf{x}-\mathbf{x}_i\|^2)$ is the RBF kernel function; γ is the RBF kernel parameter that controls the influence of each support vector; and b is the bias term.

3.6 DEA-SVR Hybrid Model

The DEA becomes too complex for a dataset having a higher-dimensional space to measure relative efficiency for frontier analysis. Even if a new DMU is added to obtain efficiency of that DMU, it is not feasible to measure the relative efficiency of the newly added DMU without re-running the DEA programme. The proposed hybrid algorithm aims to predict the efficiency of higher educational engineering institutions.

Algorithm 1: A novel hybrid algorithm of the DEA–SVM model

Input: Featured Training Dataset TFD and Testing Dataset PFD, including
Set of Featured Variables $FV(V_1, V_2, \ldots, V_p)$
Set of Featured Input Variables $FI(FI_1, FI_2, \ldots, FI_n)$
Set of Featured Input Variables $FO(FO_1, FO_2, \ldots, FO_m)$
For all DMUs $(i = 1, 2, \ldots, s)$
Output: Predicted value of Relative Efficiency (RE).

```
DEA_SVR_Predictor(trainedModel, PFD)
  Read the Featured Training Dataset TFD.
  TE_i = SolveDEA_CCR(TFD)
  TrainedModel = DEA_SVR(TFD + TE, kernel='RBF')
  PredictedEfficiency = TrainedModel(PFD)
  return PredictedEfficiency
```

A novel hybrid algorithm has been proposed that combines DEA and SVR with an RBF Kernel, as expressed in Algorithm 1. In the feature selection phase, a sample of 7428 engineering institutions was considered to determine their relative efficiency for performance assessment. A set of 15 distinct input parameters and 10 distinct output parameters is identified as a featured dataset for the proposed model.

Table 1. List of feature variables considered for the DEA–SVR model.

No.	Feature Variable	Notation	Category	Orientation	Timespan
1	Age of the Institute	*Iage*	–	Input	–
2	Student Intake	*Ni*	Teaching-Learning	Input	Previous academic year
3	Student Strength	*Ns*	Teaching-Learning	Input	Previous academic year
4	Students per faculty	*Spf*	Teaching-Learning	Input	Previous academic year
5	No. of PhD enrolled	*Nphd*	Teaching-Learning	Input	Previous academic year
6	No. of faculty members	*Ft*	Teaching-Learning	Input	Previous academic year
7	No. of faculty members having Ph.D. qualifications	*Fphd*	Teaching-Learning	Input	Previous academic year
8	No. of young faculty (having less than 8 years' experience)	*Fyoung*	Teaching-Learning	Input	Previous academic year
9	No. of experienced faculty (having at least 8 years' experience)	*Fexp*	Teaching-Learning	Input	Previous academic year
10	Capital expenditure per student (in INR)	*Ecap*	Financial Resources	Input	Previous three financial years
11	Operational expenditure per student (in INR)	*Eoper*	Financial Resources	Input	Previous three financial years
12	No. of students admitted from other states	*Nos*	Teaching-Learning	Input	Previous academic year
13	No. of students admitted from other countries	*Noc*	Teaching-Learning	Input	Previous academic year

Table 1 (continued): List of feature variables considered for the DEA–SVR model.

No.	Feature Variable	Notation	Category	Orientation	Timespan
14	No. of women students	*Nw*	Teaching-Learning	Input	Previous academic year
15	No. of women faculty members	*Fw*	Teaching-Learning	Input	Previous academic year
16	No. of publications reported in Web of Science database	*Pwos*	Research Outcome	Output	Previous three calendar years
17	No. of citations received against publication in Web of Science database	*Cwos*	Research Outcome	Output	Previous three calendar years
18	No. of publications reported in Scopus database	*Pscopus*	Research Outcome	Output	Previous three calendar years
19	No. of citations received against publication in Scopus database	*Cscopus*	Research Outcome	Output	Previous three calendar years
20	Sponsored research fundings received per faculty member	*RFpf*	Research Outcome	Output	Previous three financial years
21	Percentage of placed students in campus placement	*Nplaced*	Research Outcome	Output	Previous three academic years
22	Percentage of students opted higher studies	*Nopted*	Research Outcome	Output	Previous three academic years
23	Median of salary offered to the students in campus placement	*MSplaced*	Graduation Outcome	Output	Previous three academic years
24	No. of full-time PhD Scholars graduated	*Ngphd*	Graduation Outcome	Output	Previous three academic years
25	Percentage of graduated students	*Ngue*	Graduation Outcome	Output	Previous three academic years

Implied on the DEA-CCR model to determine the relative frequency among all considered higher educational engineering institutions. During the standardisation phase, all input-oriented features are normalised using the standard z-score normalization technique as described in Equation (3) and all output-oriented features are normalised using Min–Max normalisation as mentioned in Equation (4).

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{3}$$

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{4}$$

The featured dataset has been split into 90% for training and 10% for testing. In the DEA phase, actual efficiency has been calculated using the DEA-CCR model on a training dataset provided with multi-input and multi-output features. The actual relative efficiency is measured for each engineering institution in the training dataset. Thereafter, in the machine learning phase, the Support Vector Regression (SVR) model is trained using an RBF kernel, with the training feature datasets and the calculated actual efficiency of each institution. A test dataset has been supplied to the trained DEA-SVR model, which returns predicted efficiency values for each engineering institution.

4. RESULTS AND DISCUSSIONS

The proposed model achieved the predicted efficiency index of each engineering institution considered, testing the featured dataset. The detailed results and their discussion, along with the residual analysis, are presented below.

4.1 Confusion Matrix

The actual and predicted efficiency indices are continuous and can be categorised into class labels. Both efficiency indices are categorised into three classes: "Outstanding" for efficiency greater than 0.75, "Meets Improvement" for efficiency between 0.50 and 0.75, and "Needs Improvement" for efficiency less than 0.5. The confusion matrix for the actual efficiency category v/s predicted efficiency category is shown in Figure 2.

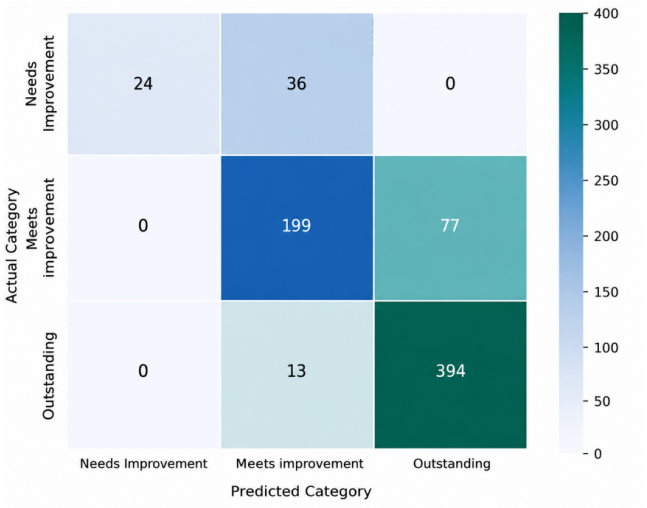


Figure 2. Confusion matrix for actual efficiency category v/s predicted efficiency category.

As depicted in Figure 2, True Positive (TP) is 683 HEIs, i.e. 394 for "Outstanding", 199 for "Meets Improvement" and 24

for "Needs Improvement" category. True Negative (TN) is 0 since none of the HEIs has predicted category as "Needs Improvement" for which HEIs have "Needs Improvement" actual category. A total of 36 HEIs obtained False Positive (FP) results, i.e. 36 HEIs have "Need Improvement" actual category, which obtained predicted category as "Meets Improvement". Important precision measures are calculated to assess accuracy and precision, as shown in Table 2.

Table 2. Precision measures of the confusion matrix for actual efficiency category v/s predicted efficiency category.

Measure	Formula	Value
True Positive	TP	683
False Positive	FP	36
False Negative	FN	0
True Negative	TN	24
Sensitivity	$TP/(TP+FN)$	1.00
Specificity	$TN/(FP+TN)$	0.40
Precision	$TP/(TP+FP)$	0.95
Negative Predictive Value	$TN/(TN+FN)$	0.96
False Positive Rate	$FP/(FP+TN)$	0.60
False Disc. Rate	$FP/(FP+TP)$	0.05
False Neg. Rate	$FN/(FN+TP)$	0.00
Accuracy	$(TP+TN)/(P+N)$	0.95
F1 Score	$2TP/(2TP+FP+FN)$	0.97

4.2 Residual Analysis for Actual Efficiency v/s Predicted Efficiency

Residual Analysis provides a graphical representation of the differences between the actual and predicted values of the target variable. Figure 3 shows the actual and predicted efficiencies of 743 engineering institutions in the testing dataset. It also indicates that the predicted efficiency lies on the same scale as the actual efficiency plotted.

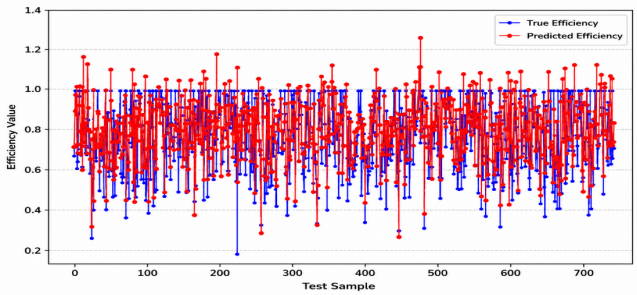


Figure 3. Graph representation of actual (true) efficiency and predicted efficiency.

Table 3 presents the residual measures used to calculate Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) based on actual efficiency y_i and predicted efficiency $\hat{y}_i$. MAE represents the average deviation between actual and predicted efficiency, as expressed in Equation (5). Since the MAE of the prediction model is 0.0703, indicating an accurate model, the prediction of the efficiency index has been achieved. Similarly, MSE represents the squared difference between actual and predicted efficiency, as expressed in Equation (6). The MSE of the prediction model is 0.0080, indicating a lower standard error. Thereafter, RMSE is the average of the squared differences between actual and predicted efficiency, as shown in Equation (7). The RMSE is 0.0894, indicating an accurate model for predicting HEI efficiency.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{8}$$

The R2 score is a numerical indicator of how well the target variables are explained by the independent variables. It also explained the model's goodness of fit based on the variability occurring in predicted variables due to the considered independent features. The R2 score is 0.7642, indicating that the model accounts for 76.42% of the variability in predicted efficiency across the considered multi-input and multi-output indicators of engineering HEIs.

Table 3. Residual measures of the confusion matrix for actual efficiency v/s predicted efficiency.

Measure	Formula	Value
Mean Absolute Error (MAE)	Eq. (5)	0.0703
Mean Squared Error (MSE)	Eq. (6)	0.0080
Root Mean Squared Error (RMSE)	Eq. (7)	0.0894
R2 Score	Eq. (8)	0.7642
Shapiro–Wilk Test	W-statistic	0.9644
Shapiro–Wilk Test	p-value	0.0000

The histogram of residuals between actual and predicted efficiency is shown in Figure 4. It shows that the maximum number of DMUs (i.e., HEIs) lie within the residual range of -0.1 to 0.1. It can be depicted that residuals follow a normal distribution as per the bell-shaped curve of residuals in the histogram. The Quantile–Quantile (Q–Q) plot indicates the scatter diagram between sample quantiles and theoretical quantiles of the featured variable, which is used to check normality of a featured variable. The Q–Q plot of residuals shows that residuals are normally distributed since the maximum residual points are residing on the reference line. The Shapiro–Wilk test is used to assess the normality of a feature variable. The p-value of the Shapiro–Wilk test is almost zero at the 5% significance level, with a test statistic of 0.9644, as shown in Table 3. Hence, the residuals of the predictive model are normally distributed. The measures are significant for the predicted model's efficiency in predicting engineering HEIs, with higher accuracy and a higher R-squared score.

5. CONCLUSION

The efficiency prediction model is developed for the performance assessment of engineering higher educational institutions using a combined algorithm of data envelopment analysis and supervised machine learning support vector regression. A model is trained on a 7432 engineering HEIs

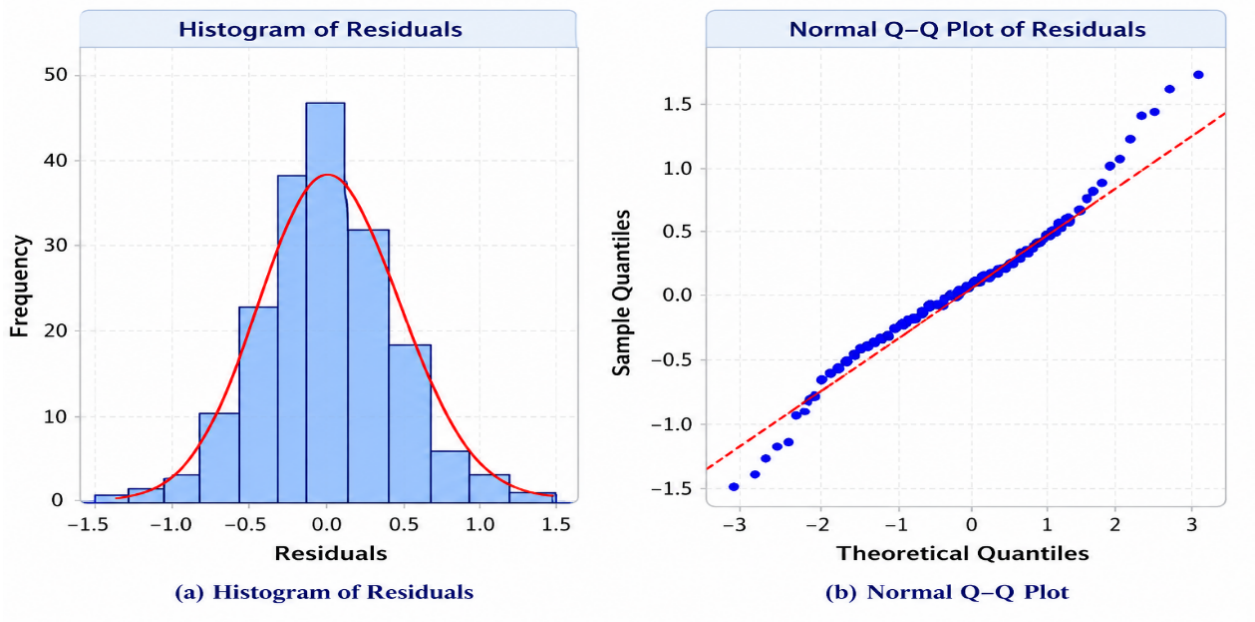


Figure 4. Histogram and Q–Q plot of residuals for actual efficiency and predicted efficiency.

dataset, with 15 multi-input features and 10 multi-output features. The predicted and actual efficiency indices of HEIs are grouped into three labels: "Needs Improvement", "Meets Expectations", and "Outstanding".

Based on the classification, the confusion matrix comprises TP, FP, FN, and TN, with the numbers of HEIs 683, 36, 0, and 24, respectively. A predictive model is derived with higher precision and an accuracy of 0.95, and an F1-Score of 0.97. The efficiency prediction model also yielded MAE, MSE, and RMSE values of 0.0703, 0.0080, and 0.0894, respectively. It indicates that the best-fitting prediction model is achieved for estimating the efficiency of HEIs in engineering. It is also noted that the R-squared score of the prediction model is 0.7642, i.e., 76.42% of the variation in predicted efficiency is explained by the independent feature variables in the training dataset.

In addition, the residuals of the prediction model are normally distributed at the 5% level of significance, with a test statistic value of 0.9644. A predictive model can also be generalised using a hybrid approach that combines DEA and SVR across multiple sectors, such as agriculture, banking, and transportation. The predictive model is helpful to government, funding agencies, and regulatory bodies in allocating resources for teaching and learning, whether financial or non-financial, and whether academic or research-oriented.

REFERENCES

[1] Department of Higher Education, Ministry of Education, Government of India, "All india survey on higher education 2021–22," Government of India, New Delhi, Tech. Rep., 2023.

[2] Ministry of Human Resource Development, Government of India, "National education policy 2020," Government of India, New Delhi, Tech. Rep., 2020.

[3] National Assessment and Accreditation Council, "Assessment and accreditation," accessed: Nov. 18, 2024.

[4] National Institutional Ranking Framework, "Parameters," accessed: Dec. 31, 2025.

[5] Department of Higher Education, Ministry of Education, Government of India, "India rankings 2024," National Institutional Ranking Framework, New Delhi, Tech. Rep., Aug. 2024.

[6] Y. Cheng, J. Peng, Z. Zhou, X. Gu, and W. Liu, "A hybrid DEA-AdaBoost model in supplier selection for fuzzy variable and multiple objectives," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 12 255–12 260, Jul. 2017.

[7] Z. Zhang, Y. Xiao, and H. Niu, "DEA and machine learning for performance prediction," *Mathematics*, vol. 10, no. 10, p. 1776, May 2022.

[8] N. Zhu, C. Zhu, and A. Emrouznejad, "A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of chinese manufacturing listed companies," *Journal of Management Science and Engineering*, vol. 6, no. 4, pp. 435–448, Dec. 2021.

[9] Y. Khoubrane, N. A. Ramli, and S. S. M. Khairi, "Performance measurement: Machine learning as a complement to DEA for continuous efficiency estimation," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 20, no. 2, pp. 288–301, Mar. 2024.

[10] P. Luangpaiboon, C. Phinkrathok, W. Atthirawong, and P. Aungkulanon, "Driving educational excellence: A data envelopment analysis study for decision-making enhancement," *SAGE Open*, vol. 14, no. 2, p. 21582440241261008, Jun. 2024.

[11] M. G. Perroni, C. P. da Veiga, E. Forteski, D. A. B. Marconatto, W. V. da Silva, C. O. Senff, and Z. Su, "Integrating relative efficiency models with machine learning algorithms for performance prediction," *SAGE Open*, vol. 14, no. 2, p. 21582440241257800, Jun. 2024.

[12] S. Yang, Y. Ogawa, K. Ikeuchi, R. Shibasaki, and Y. Okuma, "Post-hazard supply chain disruption: Predicting firm-level sales using graph neural network," *International Journal of Disaster Risk Reduction*, vol. 110, p. 104664, Aug. 2024.

[13] J. L. Zofio, J. Aparicio, J. Barbero, and J. M. Zabala-Iturriagagoitia, "Benchmarking performance through efficiency analysis trees: Improvement strategies for colombian higher education institutions," *Socio-Economic Planning Sciences*, vol. 92, p. 101845, Apr. 2024.

[14] Y. Shi and W. Zhao, "An integrated machine learning and DEA-predefined performance outcome prediction framework with high-dimensional imbalanced data," *INFOR: Information Systems and Operational Research*, vol. 62, no. 1, pp. 100–129, 2024.

[15] A. R. Dipierro and K. De Witte, "The underlying signals of efficiency in european universities: A combined efficiency and machine learning approach," *Studies in Higher Education*, vol. 50, no. 6, pp. 1306–1325, Jun. 2025.

[16] Implementation Core Committee, National Institutional Ranking Framework, "India rankings 2024: Methodology for ranking of academic institutions in india—ranking metrics for engineering," Ministry of Education, Government of India, New Delhi, Tech. Rep., 2024.

[17] J. P. S. Joorel, A. Kumar, H. Solanki, V. Raja, D. Shah, P. Pradhan, K. Trivedi, and P. Singh, "Five years of india rankings (2016–2020): An evolutionary study," *DESIDOC Journal of Library and Information Technology*, vol. 41, no. 1, pp. 42–48, Jan. 2021.

[18] Clarivate, "Web of science," citation database; accessed Nov. 18, 2024.

[19] Elsevier, "Scopus," abstract and citation database; accessed Nov. 18, 2024.