A Comparative Benchmark of Machine Learning Models for Predicting University Rankings: Insights for Higher Education Analytics

Sherin Hijazi[1], Mahmoud Moshref [2]

[12]Palestine Technical University - Kadoorie, Department of Computer Science, Palestine, Tulkarem

[2]AL-Zaytoonah University of Jordan, Department of Computer Science, Jordan, Amman

E-mail: [1]s.hijazi@ptuk.edu.ps, [2]m.moshref@zuj.edu.jo

Corresponding Author: Sherin Hijazi, s.hijazi@ptuk.edu.ps

HOW TO CITE:

Sherin Hijazi, Mahmoud Moshref (2026). A Comparative Benchmark of Machine Learning Models for Predicting University Rankings: Insights for Higher Education Analytics. International Journal of Special Education, 41(5s), 628-640.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors. Open access publication underthe terms and conditions of the Creative CommonsAttribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ABSTRACT:

QS World University Rankings are so important for evaluating academic performance and guiding strategic decision-making. Though great strides have been made to address this using machine learning and deep learning techniques, the majority of prior works suffer from outdated data use, lack of complete comparisons between state-of-the-art algorithms, and a lack of feature selection procedures that can cause bias in results beneficial towards a specific algorithm. This paper aims to fill these gaps and provide an integrated methodological approach. The analysis was based on the latest QS data (2023 and 2024) from 1,498 institutions and 26 performance indicators. After the data cleansing process, a multi-criteria weighting procedure (Information Gain, Correlation Coefficient, Relief algorithm and Gini Index) was used to evaluate feature importance with the aim of only selecting attributes inside each cross-validation training loop based on heuristic rules (the PART approach) so as not to be faced with performance penalties. Next, a total of six different predictive models were constructed and tested: (1) Linear Regression; (2) Decision Trees; (3) Random Forests; (4) Gradient Boosting Trees; (5) Multi-Layer Perceptrons and finally, (6), Convolutional Neural Network. The overall results showed that both Gradient Boosting Trees and Convolutional Neural Network models achieved the best performance among the evaluated models under the current experimental setup with the highest R-squared (0.98) together with their predictive accuracy 97.8% and lowest Mean Squared Error of 0.06 respectively. Feature importance analysis indicated that Academic Reputation and Research Impact are the most important features. This study adds value by: (1) proposing a transparent and reproducible approach that incorporates extensive benchmarking with conservative feature selection, (2) updating the performance benchmark through recent data and (3) providing meaningful insights for evidence-based strategic decision-making in higher education institutions. However,

the findings are limited to QS datasets (2023–2024) and may not generalize to other ranking systems or contexts.

Keywords: World University Rankings; Machine Learning; Deep Learning; QS Ranking Prediction; Feature Selection; Selection Bias; PART Approach.

1. Introduction

The QS World University Rankings are considered a primary means for evaluating an institution's excellence and competitiveness globally; (Chen & Chan, 2021). The QS World University Rankings is based on a variety of factors that are designed to evaluate academic reputation, employer reputation, faculty-to-student ratio, research impact and the level of internationalization within an institution (QS Intelligence Unit, 2024). Therefore, this has led to increased competition among universities in which they use global university rankings to assist them with developing policy and enhancing their status (Polyakov et al., 2020).

Learning analytics and predictive modeling have provided new avenues of research for identifying ways to understand and improve educational results. It is demonstrated through research that machine learning techniques can be applied to large amounts of data regarding students and institutions. Aunimo et al. (2025) successfully used data from learning management systems (LMS) along with student registration records to identify various elements which influence study time. They were able to demonstrate the potential of using institutional data to produce predictions in learning analytics. Similar to this example, Liu et al. (2025) showed how contrastive learning could be used to increase fairness within predictive analytics models related to online educational systems. Thus, they demonstrated the growing presence of ethical considerations when developing AI applications related to education. Both examples illustrate the growth in the application of advanced computational methodologies to higher education issues such as predicting or analyzing university rankings.

Recent advances in data analytics have enabled the application of machine learning and deep learning approaches for the investigation of intricate patterns in ranking data (Leydesdorff et al., 2021), regression models (Liu et al., 2022), decision trees (Doshi et al., 2021), ensemble techniques (Basireddy et al., 2024) and hybrid learning approaches (Kareem & Abdullah, 2025) which show the promising capability for enhancing predictive performance.

Despite this, a literature review suggests three related methodological gaps that undermine the accumulation of knowledge and its practical reliability in these subject areas: First there is the temporal data gap which occurs when many models are based on historical data that does not reflect actual current trends (Udupi et al., 2023). Second, the methodological comparison gap, a few studies offered a comprehensive benchmark for various advanced machine learning and deep learning algorithms under a single experimental condition (Wisaeng & Muangmeesri, 2025). Third, there is a transparency gap in feature selection, where applied studies often neglect to document the feature selection approach or follow the traditional practice of using all data before partitioning (the ALL approach), which exposes the results to the risk of feature set selection bias and reduces the possibility of replication (Aldehim & Wang, 2017).

In this paper presents an integrated methodological framework to fill these gaps. It stands out due to three main contributions: (1) the use of the most recent QS ranking data (2023- 2024); (2) a thorough benchmarking comparison between six different predictive models evaluated in a standardized experimental environment; and (3) the systematic procedure followed with a multi-criteria weighting method for feature selection, respecting strictly evaluation and selection only through training data while using cross-validation cycle (the PART approach). This provides a transparent and reproducible framework, proactively reducing the risk of bias. Furthermore, this improved

methodology demonstrated strong predictive performance and yielded valuable interpretive insights into the factors that determine ranking.

The remainder of the paper is organized as follows: Section 2 reviews literature review. Section 3 presents the proposed methodology and experimental design. Section 4 discusses experimental results. Section 5 conclusion and future works.

2. Literature Review

In this section, we summarize the existing research of predicting university rankings focusing on machine learning and deep learning approaches.

2.1 Predictive Studies Using Traditional Machine Learning

Several studies have investigated that have reputed with traditional methods especially ML, to predict university rankings (Sekhon & Oberoi, 2024; Wardley et al., 2024). The authors Udupi et al. (2023) employed automatic regression models using QS data. They found that linear models are suitable baseline methods to detect varying relations. However, they observed that such models were limiting for complex relationships. But they did not obtain suitable results when modeling complex relationships. Similarly, Liu et al. (2022) employed the decision tree as a data mining approach. They discovered that many such methods suffered from overfitting with small data. The authors Gokul and Gutjahr (2022) conducted comparisons for different ML predictions of university scores. They found that for censored regression, these techniques were indeed practical, although ensemble methods (e.g., Random Forests) worked best. Notwithstanding their importance, most of these studies were based on old data or only a small number of models, therefore being unable to offer an exhaustive benchmarking overview (Basireddy et al., 2024; Estrada-Real & Cantu-Ortiz, 2022).

2.2 Predictive Studies Using Deep Learning and Hybrid Models

As the complexity of data grows, an emerging trend in research is to investigate the power of deep learning as well as hybrid models. The authors Basireddy et al. (2024) developed hybrid models combining Support Vector Regression, Random Forest, and Linear Regression for QS ranking prediction, showing improved accuracy through model integration. In another effort, Kareem and Abdullah (2025) developed a hybrid deep learning model that combines aspects of business intelligence analytics. This demonstrated that architecture like Bi-LSTM handles complex dependencies in the data. Another study by Wisaeng and Muangmeesri (2025) developed a model that utilized a hybrid feature selection technique with several ML models. Although these studies make significant progress, they do not typically provide a systematic comparison of their findings with ML and/or DL techniques in general. Alongside such architectural novelties, a few works have investigated feature engineering beyond the classical institutional features. For instance, Abuein et al. (2021) utilized author collaboration statistics as predictor features from research networks, indicating that hybrid models can be enhanced by incorporating bibliometric network analysis in addition to traditional ranking indicators. However, these methods are relatively unexplored in the mainstream ranking prediction literature of information retrieval.

2.3 Studies on Feature Selection in the Ranking Context

In modeling high-dimensional data such as ranking, the selection and processing of features are crucial steps. The literature about filter methods e.g., information gain and correlation coefficient justify their role in dimensionality reduction and model interpretability. The influential study of Aldehim and Wang (2017) pointed out the methodological problem of directly using data for feature selection, in order to circumvent feature subset selection bias. It demonstrated that the conventional approach of using all data (the ALL approach) before splitting can lead to over-optimistic results, whereas the approach conducting selection only on training data (the PART approach) within a cross-validation cycle is more reliable for estimating true model performance, especially with smaller datasets. However, in the vast majority of applied studies it is not quite clear whether some feature selection was involved and how that step was performed, which

means for those results are not reproducible. In addition to filter and wrapper methods, the FS literature provides additional dimensionality reduction techniques. Methods such as Sparse Principal Component Analysis (Zou et al., 2006) can generate explanatory linear combinations of original features and could model synergistic effects between correlated ranking indicators. However, their use in the task of university ranking prediction is also limited, with most works still favoring simple filter methods. While alternative analytical approaches such as clustering have been proposed to reveal hidden structures and commonalities among universities that traditional linear rankings may not capture (Matlis et al., 2024).

2.4 Research Gap and Distinction

The review of the literature shows that there are three interrelated methodological gaps, as presented in Tables 1 and 2. These are: a temporal data gap, since models are trained on historical data and do not reflect current changes in ranking methods (Udupi et al., 2023). A gap in methodological comparison, as a small number of works offering an extensive benchmarking comparison of different advanced ML and DL approaches on the same data using the common experimental methodology (Wisaeng & Muangmeesri, 2025). Lack of transparency in feature selection, lack of clearly documented and systematic procedures for feature selection (Aldehim & Wang, 2017).

These gaps are the focus of this research directly. Our contributions are: (1) use of the latest QS ranking data (2023 & 2024); (2) direct benchmarking and comparison of six diverse predictive models in a consistent experimental set-up (3) systematic implementation and reporting of an aggregated-multi-criteria-weight based feature selection scheme, conforming to the principle of performing selection within the CV cycle (PART approach), not to suffer from selection bias, thus delivering a transparent-and-reproducible methodological framework.

Table 1: Comparison of Previous Studies Models

Study	Data Source	Key Features	Models Used
Basireddy et al. (2024)	QS World University Rankings	Not Specified	Hybrid (SVR, RF, LR)
Kareem and Abdullah (2025)	QS + Multi-dimensional Data	Multi-dimensional	Hybrid DL (Bi-LSTM)
Udupi et al. (2023)	QS Data	Several variables	Auto-regression, Linear Models
Wisaeng and Muangmeesri (2025)	QS Data	Not Specified	Various ML Models
Gokul and Gutjahr (2022)	University Scores	Multiple	Censored Regression, RF, others
Proposed Study	**QS World Rankings 2023-2024 (N=1,498)**	**26 Performance Indicators**	**LR, DT, RF, GBT, MLP, CNN**

Table 2: Comparison of Previous Studies Gap analysis

Study	Feature Selection Strategy	Key Limitations / Gaps
Basireddy et al. (2024)	Not Specified	Data specifics and feature selection process are not transparent.
Kareem and Abdullah (2025)	Not Specified	Lacks benchmarking against a broad set of ML/DL baselines.

Udupi et al. (2023)	Not Specified	Dataset size/recency unclear; limited to linear models.
Wisaeng and Muangmeesri (2025)	Hybrid Selection	Model comparison is limited; data size and recency not specified.
Gokul and Gutjahr (2022)	Not Specified	Does not explore deep learning; feature engineering not detailed.
Proposed Study	**Multi-Criteria Weighting with PART application**	**--**

However, many of these studies lack reproducible experimental settings.

2. Methodology

This section presents the overall research framework employed to develop and evaluate predictive models for the QS World University Rankings. The approach is organized into 4 steps, based on below stages: (1) Data collection and preprocessing, (2) Feature selection, (3) Model development, and (4) Experimental set-up. Figure 1 provides the complete steps comprising the overall research framework, employed in this research and results validation.

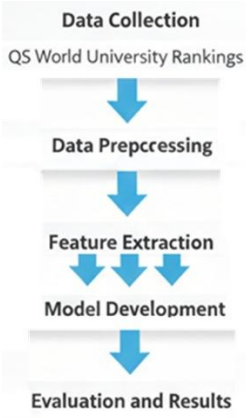


Figure 1: Overall Research Framework

3.1. Dataset Description and Preprocessing

3.1.1. Data Source and Collection

The study utilized 2023 and 2024 QS World University Rankings data available on Kaggle. The sample covered 1,498 universities from around the world and across 26 indicators. These indicators also correspond with the six key components of the QS methodology: academic reputation, employer reputation, faculty-student ratio, citations per faculty, proportion of international faculty and proportion of international students.

3.1.2. Data Cleaning and Imputation

Missing values for all numerical features, missing values were detected in this study. Since missingness in these features was minimal and as we wanted to preserve the entire dataset for precision analysis, we filled missing data by average value method among these variables. This is an easy and stable estimator (Udupi et al., 2023). Mean imputation may introduce bias when data is not missing completely at random.

3.1.3. Outlier Detection and Handling

Outliers-led noise was detected and down-weighted using the k-nearest neighbors' algorithm (k = 10) approach deployed in the study to ensure accuracy of data. Individual data points with large average distances from neighboring ones were located. Instead of removing these potential outliers, they were adjusted (to the 99-th percentile) while saving their information and reducing their negative influence on the model training. Outlier adjustment may affect extreme-case interpretability.

3.1.4. Data Normalization

Z-score standardization was used in the study so that all features could contribute to the model's learning, based on their original metrics. This converted all numerical features to a mean of zero and a standard deviation of one, which is particularly useful for gradient-based optimization algorithms used in neural networks and for distance-based models.

3.1.5. Data Splitting Strategy

The study divided the pre-processed dataset into a 70% training set and a 30% test set. Stratified sampling based on the target variable (ranking score decile) was used to ensure that the distribution of university rankings was proportional in both subsets, thus enhancing the model's ability to generalize to unseen data.

3.2. Feature Selection

The study employed a weighted feature selection methodology to identify the most influential features in the university production process. These weights include:

Weighting by Information Gain (W. Info.): Information gain measures the reduction in entropy when a dataset is partitioned based on a specific feature. The goal is to identify the feature that provides the most useful information for predicting the target variable.

Weighting by Gini Index (W. Gain): The Gini index is a measure of the heterogeneity or diversity of a dataset. It determines how often a randomly selected item from the dataset is misclassified if randomly assigned to the distribution of classifications within the subset.

Weighting by Relief (W. Relief): This assesses the quality of a feature by repeatedly sampling it and comparing it to its nearest neighbors. The core idea is to evaluate how well a feature separates samples from different categories (while keeping samples within the same category closely related).

Weighting by Correlation (W. Corr.): This is a feature selection technique where features are assigned weights based on their correlation with the target variable (category). This technique identifies features that are strongly related to the target variable and can help reduce dimensions by selecting the features most relevant to predictive modeling.

3.3. Predictive Models

The investigation used and compared many of the six most common predictive models such as classical linear models, ensemble methods, and deep learning architectures.

Linear regression (LR): A basic model in linear relations.

Decision tree (DT): A natural and interpretable model for nonlinear decision boundary.

Random Forest (RF): An ensemble of decision trees with clustering aiming to minimize variance, and overlap, thereby demonstrating a robust performance.

Gradient Boosted Trees (GBT): Effective clustering approach, where the trees are built in a stage-wise fashion and each new tree corrects the errors of previous trees which eventually leads to high predictive performance (Basireddy et al., 2024).

Multilayer Perceptron (MLP): A common type of front-end neural network consisting of multi hidden-layers which can learn complex non-linear patterns using ReLU activation functions.

Convolutional Neural Network (CNN): Adjusted to the one-dimensional organization of feature data. The 1D-CNN structure is formulated to apply convolutional filters over the feature vector and area to enable the model in identifying local patterns, as well as hierarchal interactions among related ranking indexes (Kareem & Abdullah, 2025).

3.4. Experimental Setup and Evaluation

3.4.1. Model Training and Hyperparameter Tuning

In the present study, all models were trained with a training set of 70 %. The hyper- parameters of the complex models (RF, GBT, MLP, CNN) were tuned by grid search over a fivefold cross-validation on the training data in order to avoid information leakage from test set. The relevant hyperparameters that were optimized in our experiments are tree depth, number of estimators, learning rates and the properties of the network architecture (number of layers, number of neurons and dropout). Hyperparameter ranges were selected empirically based on prior studies.

3.4.2. Performance Evaluation Metrics

The best models in the end were tested on a 30% held-out set making use of a large range of metrics to test various aspects of performance:

R-squared (R^2): Quantifies the percentage of variance in the target that is explained by the model.

Mean Squared Error (MSE): Measures the average of the squares of the errors between Actual and Predicted values.

Accuracy: It represents the proportion of predictions for which the predicted rank ±3 ranks is close to the true value, offering a comfortable way to measure classification performance.

3.4.3. Computational Implementation and Feature Analysis

All experiments were conducted on a machine with 8 GB of RAM and an Intel Core i5 CPU with speed of 2.4 GHz. The models were developed by Python (3.10.8) language, utilizing RapidMiner (v10.3) as an analytical instrument to help build, train, and test the model. and using libraries such as Scikit-learn for conventional ML models and TensorFlow / Keras for deep learning models. The code and environment settings are posted for complete reproducibility.

Weight-based feature selection to make the model more efficient and simpler, features were selected based on their weight to use only those most important in the process of classifying universities, that would produce more accurate and efficient predication models. Subsequently, a correlation matrix was used to investigate the associations among them before-and-after selection with emphasis on identifying redundant variables and mitigating multicollinearity (Aldehim & Wang, 2017). In the Table 3 contains four weight matrices (W. Info.), (W. Corr.), (W. Relief), and (W. Gini).

Table 3. Weight-Based Feature Importance Scores for the Combined QS 2023-2024 Dataset

Feature	QS Ranking 2023-2024 Dataset with Error ±5%				
	W. Info.	W. Corr.	W. Relief	W. Gini	Avg. Weight
Academic Reputation Rank	1.00 ± 5%	1.00 ± 5%	1.00 ± 5%	1.00 ± 5%	1.00 ± 5%
Academic Reputation Score	1.00 ± 5%	0.90 ± 5%	0.49 ± 5%	1.00 ± 5%	0.85 ± 5%
Age Band	0.07 ± 5%	0.27 ± 5%	0.06 ± 5%	0.08 ± 5%	0.12 ± 5%
Citations per Faculty Rank	0.62 ± 5%	0.89 ± 5%	0.64 ± 5%	0.65 ± 5%	0.70 ± 5%
Citations per Faculty Score	0.61 ± 5%	0.88 ± 5%	0.51 ± 5%	0.65 ± 5%	0.66 ± 5%
Employer Reputation Rank	0.53 ± 5%	0.78 ± 5%	0.59 ± 5%	0.56 ± 5%	0.62 ± 5%
Employer Reputation Score	0.53 ± 5%	0.77 ± 5%	0.35 ± 5%	0.56 ± 5%	0.55 ± 5%
Employment Outcomes Rank	0.46 ± 5%	0.75 ± 5%	0.42 ± 5%	0.49 ± 5%	0.53 ± 5%
Employment Outcomes Score	0.46 ± 5%	0.66 ± 5%	0.13 ± 5%	0.49 ± 5%	0.44 ± 5%
Faculty Student Rank	0.52 ± 5%	0.78 ± 5%	0.89 ± 5%	0.50 ± 5%	0.67 ± 5%
Faculty Student Score	0.52 ± 5%	0.79 ± 5%	0.71 ± 5%	0.50 ± 5%	0.63 ± 515
Focus	0.10 ± 5%	0.28 ± 5%	0.19 ± 5%	0.11 ± 5%	0.17 ± 5%
International Faculty Rank	0.42 ± 5%	0.69 ± 5%	0.39 ± 5%	0.44 ± 5%	0.49 ± 5%
International Faculty Score	0.42 ± 5%	0.70 ± 5%	0.45 ± 5%	0.44 ± 5%	0.50 ± 5%
International Research Network Rank	0.56 ± 5%	0.84 ± 5%	0.24 ± 5%	0.60 ± 5%	0.56 ± 5%
International Research Network Score	0.56 ± 5%	0.83 ± 5%	0.23 ± 5%	0.60 ± 5%	0.56 ± 5%
International Students Rank	0.38 ± 5%	0.67 ± 5%	0.26 ± 5%	0.42 ± 5%	0.43 ± 5%
International Students Score	0.40 ± 5%	0.65 ± 5%	0.25 ± 5%	0.43 ± 5%	0.43 ± 5%
Overall Score	0.78 ± 5%	0.01 ± 5%	0.42 ± 5%	0.73 ± 5%	0.48 ± 5%
Research	0.39 ± 5%	0.64 ± 5%	0.16 ± 5%	0.42 ± 5%	0.40 ± 5%
Size	0.00 ± 5%	0.00 ± 5%	0.11 ± 5%	0.00 ± 5%	0.03 ± 5%
Status	0.02 ± 5%	0.09 ± 5%	0.00 ± 5%	0.02 ± 5%	0.04 ± 5%

Sustainability Rank	0.67 ± 5%	0.90 ± 5%	0.47 ± 5%	0.69 ± 5%	0.69 ± 5%
Sustainability Score	0.67 ± 5%	0.89 ± 5%	0.44 ± 5%	0.69 ± 5%	0.67 ± 5%

It can be seen from the table that academic reputation rank and academic reputation score have higher weights, and their average weights are 1.00 and 0.85 respectively (Table 3). The ratio of citations per faculty and sustainability measures also had an impact. In contrast, size (average weight: 0.03), status (0·04) and age band (0·12) had little predictive value. According to this quantitative ranking, low-impact features (those with average weights less than a given threshold) were not included in the subsequent model training.

The results of feature important according to the weight-based selection are illustrated in Figure 2. The results reveal that features associated with academic reputation are the leading predictors, especially Academic Reputation Rank and Academic Reputation Score. Citation analysis and the sustainability-related indicators also play a significant role. In contrast, variables such as Size, Status, Age Band, and Focus demonstrate limited influence on the model outcomes. These results confirm the usefulness of the weighting method in identifying influential features and provide a basis for excluding low-impact variables before model training.

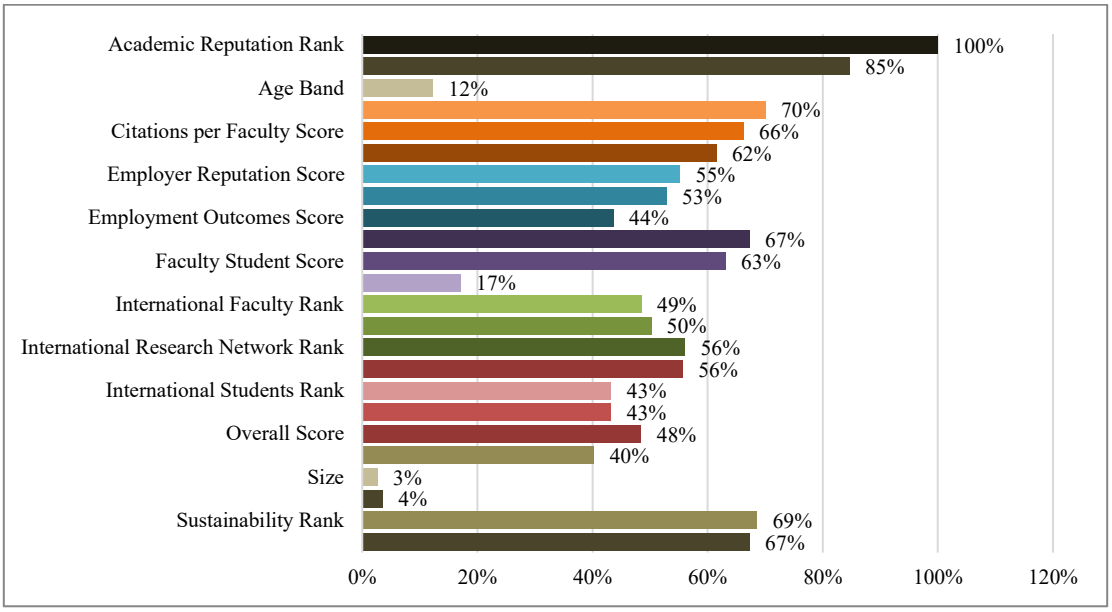


Figure 2: The percentage Influence of each feature to the prediction

Figures 3 and 4 provide correlation matrices that illustrate the relationship among characteristics and their relationship to the target variable before the feature selection process. After the feature selection process is applied, we see correlations among characteristics and the target variable as depicted in figures 5 and 6.

QS Ranking 2023 and QS Ranking 2024 datasets show considerable correlations among various reputation- related indicators, citation-based metrics and sustainability indicators, as well as these indicator's relationship to the target variable. The reputation and citations have moderate to strong correlations with the total ranking score; however, a few institutional attributes appear to have weak or no relationship to the ranking score. It also appears there are some levels of inter-correlation among some of the individual features; thus, this suggests redundancy and potential multicollinearity in the original set of features.

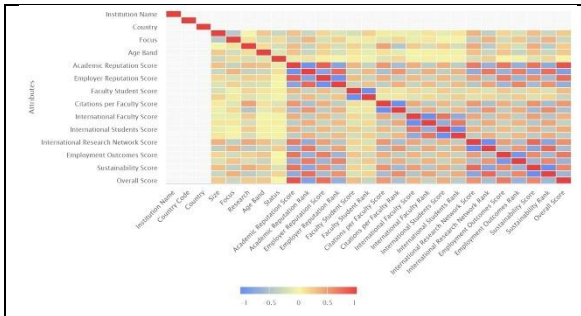


Figure 3: Feature Correlation Matrix for QS Ranging 2023 before Feature Selection

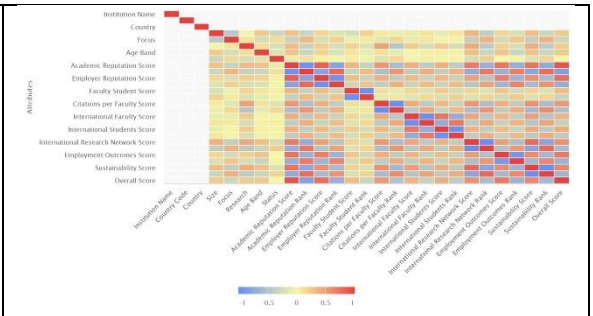


Figure 4: Feature Correlation Matrix for QS Ranging 2024 before Feature Selection

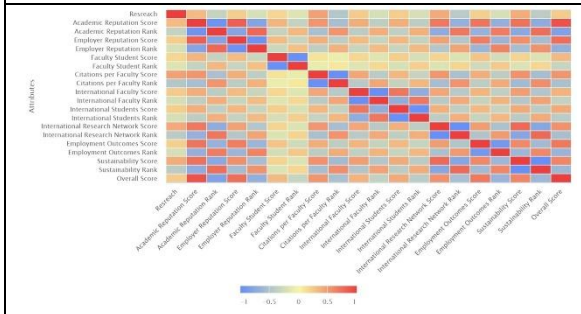


Figure 5: Feature Correlation Matrix for QS Ranging 2023 after Feature Selection

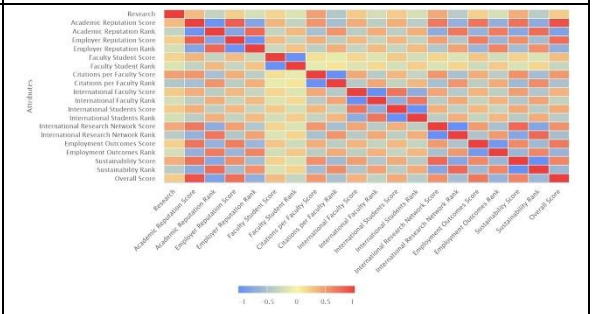


Figure 6: Feature Correlation Matrix for QS Ranging 2024 after Feature Selection

The correlation matrices from the same data sets prior to the application of feature selection; versus post-application demonstrate that feature selection has generated a subset of features with a structure and relationship (correlation) much more interpretable than those of the original features. In addition, the relationship (i.e., correlation) between the selected influential features and the target variable have been preserved. While at the same time many of the difficult-to-interpret correlations among the individual features were diminished or eliminated. Thus, the feature selection process identified many low-contribution and redundant features and removed them resulting in a smaller number of features with little multicollinearity and clearer relationships to the prediction. Overall, both correlation matrices illustrate that our feature selection methodology preserved the major features and improved feature independence as needed to achieve good quality, robust predictive models.

4. Results and discussion

4.1. Comparative Model Performance

Results for the predictive model performances on both the QS 2023 Test Set (Table 4) and the QS 2024 Test Set (Table 5) are shown. Overall, there were considerable differences in predictive model performance as well as significant difference in performance by each model compared to its respective baseline. A graphical representation of how each model compares across all evaluation criteria is found in Figures 7 and 8.

Table 4: Performance Evaluation of Predictive Models on the QS 2023 Dataset

Prediction Model	R^2	MSE	Accuracy
LR	0.90	0.13	90.5%
DT	0.91	0.13	91.3%
RF	0.95	0.09	92.2%
MLP	0.96	0.09	96.0%
GBT	0.96	0.08	93.9%
CNN	0.96	0.08	94.8%

The 2023 results establish a baseline where GBT and CNN already show superior MSE, while MLP achieves the highest accuracy.

Table 5: Performance Evaluation of Predictive Models on the QS 2024 Dataset

Prediction Model	R^2	MSE	Accuracy
LR	0.89	0.14	89.6%
DT	0.93	0.12	90.7%
RF	0.97	0.08	92.2%
MLP	0.97	0.08	96.8%
GBT	0.98	0.07	97.6%
CNN	0.98	0.06	97.8%

When we look at the 2024 results, a significant improvement is seen in CNN with the best accuracy (97.8%) and lowest MSE (0.06), followed by GBT.

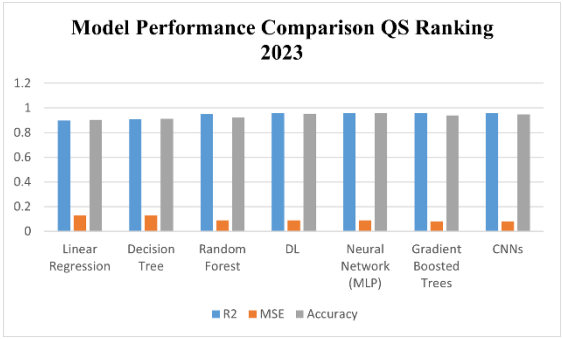


Figure 7: Model Performance Comparison QS Ranking 2023

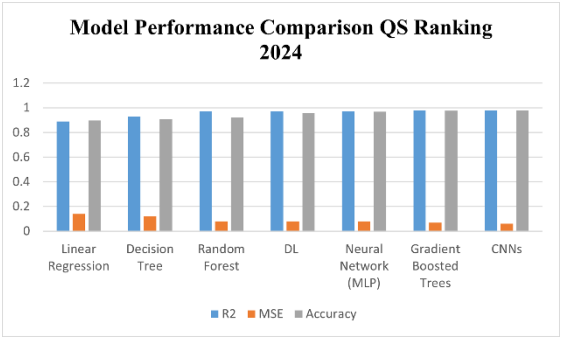


Figure 8: Model Performance Comparison QS Ranking 2024

4.2. Discussion of Results

Gradient Boosted Trees and CNN performed best on the 2024 dataset achieving R^2 values of .98 and MSE values of 0.06 while achieving an accuracy value of 97.8% suggesting they are able to model non-linear relationships within the data.

Ensemble models including RF and GBT out-performed individual decision tree models by a significant margin, indicating ensemble averages of multiple decision trees have the ability to reduce variance and increase generalizability. Although LR provided a good base line ($R^2 \sim 0.90$) it's lack of ability to saturate in regards to performance illustrates LR inability to identify complex interactions, which demonstrates that randomizing and using a deep neural network has the capability to effectively correct this deficiency.

The increased performance from 2024 results could be due to variations in data quality or model fit. Also, the exceptionally high performance of CNN, a model primarily used for spatial data, also indicates our adaptation of the architecture to be applied to tabular data identifies regional meaningful local patterns spanning the feature space.

4.3. Comparison with Previous Studies and Contributions

This paper addresses several key methodological issues that were previously addressed in the existing body of research. In particular, as demonstrated in Table 6, the approach described here represents an important advancement on prior methodologies with regard to both conceptual analysis and practical utility.

Firstly, this paper utilizes the most current QS data (2023-24) which allows for a contemporary view of the current ranking dynamics as well as methodology when compared to prior studies possibly using older or out-of-date information sets (Udupi et al., 2023).

Secondly, the results represent a complete head-to-head comparison of six different algorithms – including traditional linear regression; ensemble

models; and deep learning models – within a single unifying and replicable experimental design. The results provide a much clearer and definitive basis for comparisons among different types of models than studies which compare single or small numbers of hybrid models without comparable baseline conditions (Basireddy et al., 2024).

Thirdly, this study provides and demonstrates a multi-criteria weighting based feature selection process using an open and transparent methodology in contrast to many studies that utilize the ALL-approach (all data is used for feature selection which can lead to performance estimation bias Aldehim & Wang, 2017). The methodology presented here includes the application of a systematic process that allows feature selection to be done only through training data (PART-process) utilizing cross-validation. The combination of these processes along with the utilization of four different filter techniques (InfoGain, Correlation, Relief, GiniIndex) for creating an overall feature ranking (see Table. 3), enhances both the computational efficiency as well as the robustness of the methodologies used for modeling and interpreting study results—both elements often understated or misrepresented in predictive research studies.

The quantitative benefits from combining methodological improvements are clearly demonstrated by the best performing models. These were the GBT model and the modified version of the CNN model, each achieving a maximum R^2 of 0.98 and 97.8% accuracy on the 2024 data set as indicated by the findings reported in table six. Both of these values represent significant increases over other high value performances recorded in the literature; typical maxima for R^2 range from 0.9-0.95. Therefore, we believe this large increase represents an improvement in predictive accuracy compared to several prior studies in this area.

Therefore, we provide with this research study a robust, scaleable, reusable framework that achieves two objectives: (1) learns accurate forecasts of ranking to be used in institutional planning, and (2) provides empirical support for the underlying determinants of ranking via comprehensive feature analyses. With its combination of importance and theoretical basis, our research study contributes to a more refined tool for university decision makers and policy makers who may use such empirical evidence to make informed strategic decision.

Table 6: Comparative Analysis with Key Previous Studies

Study	Model Used	R^2	MSE	Accuracy
Basireddy et al. (2024)	Hybrid (RF + SVR + LR)	0.93	N/A	N/A
Kareem and Abdullah (2025)	Hybrid DL(Bi-LSTM)	0.95	N/A	N/A
Udupi et al. (2023)	Random Forest	0.90	N/A	N/A
Wisaeng and Muangmeesri (2025)	Hybrid (SVM + Feature Selection)	N/A	N/A	N/A
Proposed Study	**GBT on QS 2024**	**0.98**	**0.07**	**97.6%**
Proposed Study	**CNNs on QS 2024**	**0.98**	**0.06**	**97.8%**

5. Conclusion and Future Work

This article described a new, both interpretable and reproducible way to make predictions about world university rankings using recent QS data (the years 2023-2024). This was done by using conservative feature selection inside cross-validation. It was shown that the newly created non-linear models (Gradient Boosted Trees and Convolutional Neural Networks), which have strong predictive ability (Accuracy= 97.8%; R^2 = 0.98%), were significantly better than all other tested models. Therefore, it can be concluded that Academic Reputation and Research Impact are the most important factors influencing ranking results. Thus, these findings will help universities develop strategic plans.

The study has one main limitation. Data used in this study are only from QS World University Rankings (2023-2024) and therefore can reduce the applicability of results to other rankings and larger

contexts. While previous studies examined closely related issues, due to limited number of comparative studies using an equivalent integrated method, it was difficult to make direct comparisons to prior work. Future research needs to apply temporal deep learning models for long-term forecasting; multi-task learning across multiple ranking systems; integration of unstructured data and explainable AI techniques. Further advancement towards causal modeling and sub-indicator level predictions will provide additional empirical support for policy simulation and multi-objective optimization. Embedding these models into institutional decision support systems will enable predictive analytics to be transformed into proactive tools for sustainable development of higher education.

Acknowledgment

The author would like to thank the Palestine Technical University - Kadoorie for their financial support to conduct this research.

Conflict of Interest

The authors declare no competing interests and no conflict of interest.

Data Availability

The datasets analyzed during the current study are available in the kaggle repository, https://www.kaggle.com/datasets/joebeachcapital/qs-world-university-rankings-2024.

References

Abuein, Q. Q., Almahmoud, M. H., & Elayan, O. N. (2021). Improving QS rank based on the classification of authors research collaboration using machine learning techniques. In *2021 12th International Conference on Information and Communication Systems (ICICS)* (pp. 63–68). IEEE. https://doi.org/10.1109/ICICS52457.2021.9464603

Aldehim, G., Wang, W. Determining appropriate approaches for using data in feature selection. *Int. J. Mach. Learn. & Cyber.* **8**, 915–928 (2017). https://doi.org/10.1007/s13042-015-0469-8

Aunimo, L., Kauttonen, J., Vahtola, M. *et al.* Combining data from a LMS and a student register for exploring factors affecting study duration. *J Comput High Educ* **37**, 1032–1055 (2025). https://doi.org/10.1007/s12528-024-09414-4

Basireddy, C. S., Cheruku, V. K. G., Rajagopal, S., & Soangra, R. (2024). Hybrid prediction models for assessing the higher education institutions performance in QS World Institution rankings. *F1000Research, 13*, 1529. https://doi.org/10.12688/f1000research.155847.1

Chen, G., & Chan, L. (2021). University rankings and governance by metrics and algorithms. In *Research Handbook on University Rankings* (pp. 425-443). Edward Elgar Publishing.

Doshi, N., Gundam, S., & Chaudhury, B. (2021). Strategizing university rank improvement using interpretable machine learning and data visualization. *arXiv preprint arXiv:2110.09050*. https://doi.org/10.48550/arXiv.2110.09050

Estrada-Real, A. C., & Cantu-Ortiz, F. J. (2022). A data analytics approach for university competitiveness: the QS world university rankings. *International Journal of Interactive Design and Manufacturing, 16*(3), 871-891. https://doi.org/10.1007/s12008-022-00978-y

Gokul, R. P., & Gutjahr, G. (2022). Comparison of machine learning algorithms for censored regression models of university scores. *AIP Conference Proceedings, 2424*(1), 070004. https://doi.org/10.1063/5.0076867

Kareem, A. R.., & Abdullah, H. S. (2025). Business Intelligence Approach-Based Hybrid Deep Learning Model for QS World University Ranking. *Iraqi Journal of Science*, *66*(1), 375-387. https://doi.org/10.24996/ijs.2025.66.1.30

Leydesdorff, L., Wagner, C. & Zhang, L. (2021). Are University Rankings Statistically Significant? A Comparison among Chinese Universities and with the USA. *Journal of Data and Information Science*, *6*(2), 67-95. https://doi.org/10.2478/jdis-2021-0014

Liu, X., Chen, G., Wen, S., & Han, X. (2022). Analysis and prediction of QS World University Rankings based on data mining technology. In *Proceedings of the 6th International Conference on Education and Multimedia Technology* (pp. 60–64). Association for Computing Machinery. https://doi.org/10.1145/3551708.3556207

Liu, Z., Xing, W., Jiang, Y., Li, C., Kim, T., & Li, H. (2025). Leveraging contrastive learning to improve group and individual fairness in predictive analytics for online learning. *Journal of Computing in Higher Education*, *37*(4), 1341-1370. https://doi.org/10.1007/s12528-025-09468-y

Matlis, G., Dimokas, N., & Karvelis, P. (2024). Unveiling University Groupings: A Clustering Analysis for Academic Rankings. Data, 9(5), 67. https://doi.org/10.3390/data9050067

Polyakov, M., Bilozubenko, V., Korneyev, M., & Nebaba, N. (2020). Analysis of key university leadership factors based on their international rankings (QS World University Rankings and Times Higher Education). *Problems and Perspectives in Management, 18*(4), 142–152. https://doi.org/10.21511/ppm.18(4).2020.13

QS Intelligence Unit. (2024). *QS World University Rankings Methodology*. Retrieved from https://www.topuniversities.com/qs-world-university-rankings/methodology

Sekhon, A., & Oberoi, S. S. (2024). *Predictive Analysis System for National Ranking and Accreditation of HEIs*. *Communications on Applied Nonlinear Analysis*, **32**(1s), 539–551. https://doi.org/10.52783/cana.v32.2318

Udupi, P. K., Dattana, V., Netravathi, P. S., & Pandey, J. (2023). Predicting global ranking of universities across the world using machine learning regression technique. In *SHS Web of Conferences* (Vol. 156, p. 04001). EDP Sciences. https://doi.org/10.1051/shsconf/202315604001

Wardley, L. J., Rajabi, E., Amin, S. H., & Ramesh, M. (2024). A machine learning approach feature to forecast the future performance of the universities in canada. *Machine Learning with Applications*, *16*, 100548. https://doi.org/10.1016/j.mlwa.2024.100548

Wisaeng, K., & Muangmeesri, B. (2025). University rankings prediction using hybrid feature selection based on machine learning methods. *International Journal of Analysis and Applications*, *23*, 112-112. https://doi.org/10.28924/2291-8639-23-2025-112

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics, 15*(2), 265-286. https://doi.org/10.1198/106186006X113430