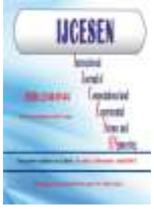
International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 11-No.1 (2025) pp. 59-70
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Towards Smarter E-Learning: Real-Time Analytics and Machine Learning for Personalized Education

N. S. Koti Mani Kumar Tirumanadham[1]*, S. Thaiyalnayaki[2], V. Ganesan[3]

[1]Research Scholar, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Selaiyur 600073, Tamil Nadu, India
* **Corresponding Author Email:** manikumar1248@gmail.com - **ORCID:** 0000-0003-3900-1900

[2]Associate Professor, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Selaiyur 600073, Tamil Nadu, India
Email: thaiyalnayaki.cse@bharathuniv.ac.in - **ORCID:** 0009-0006-4973-8147

[3]Professor, Department of Electronics and Communication Engineering, Bharath Institute of Higher Education and Research, Selaiyur 600073, Tamil Nadu, India
Email: vganesh1711@gmail.com - **ORCID:** 0000-0002-3978-8873

Article Info:

DOI: 10.22399/ijcesen.786
Received : 17 October 2024
Accepted : 23 December 2024

Keywords :

Z-score,
E-Learning,
Min-Max scaling,
Ridge regression,
Gradient Boosting Machine,
Recursive Feature Elimination.

Abstract:

E-Learning platforms change fast, and real-time behavioural analytics with machine learning provides the most powerful means to enhance learner outcomes. The datasets undergo preprocessing techniques like Z-score outlier detection, Min-Max scaling for feature normalization, and Ridge-RFE (Ridge regression and Recursive Feature Elimination) for feature selection in order to improve the accuracy and reliability of the predictions. Applying the Gradient Boosting Machine, classification accuracy up to a 94% level with respect to the model about predictions on learner outcomes was achievable. Thus, applying this, feedback systems may offer timely recommendations or directions in class that propel students toward better understanding on how to raise participation and success percentages. However, this approach has some potential benefits but there are still various challenges such as managing the data imbalance for models that generalize in a dynamic environment. Though hybrid methods mitigate this problem, real-time data pipelines with behaviour analytics incorporation call for significant computer-intensive resources and infrastructure. This integration has very high paybacks. It makes possible more responsive E-Learning platforms with individual needs almost met in real-time manners, thus giving instantaneous feedback, content suggestions, and timely interventions. Finally, convergence of real-time analytics with ML models culminates in adaptive learning environments which improve student engagement, retention, and quality of academic results.

1. Introduction

E-Learning has thus developed rapidly, using even advanced technologies like ML and AI to enhance the outcomes of learning [1]. Now, E-Learning platforms collect a huge amount of data on learner behaviour and performance, thus providing ample scope for the use of that data for personalized intervention and predictive insights. Real-time behavioural data such as login frequency, participation rate, and task completion time form the basis of integrating information for the development of an adaptive learning environment that evolves dynamically in response to student need [2].

Even with such predictive advances, class imbalances persist in datasets as an added challenge to be overcome by models to ensure good generalization in diverse educational environments. The dynamic nature of the E-Learning platform means that the student interaction evolves with time, making it difficult to measure and quantify factors such as motivation or cognitive load [3]. To that end, multimodal data like eye-tracking, facial expressions, and interaction logs, though not explored in much depth to date, would have to that end, multimodal data like eye-tracking, facial expressions, and interaction logs, though not explored in much depth to date, would have provided a much more extensive view of learner behaviour. There is also increased demand for holistic frameworks which integrate predictive models with adaptive learning strategies to create adaptive learning pathways. Such frameworks might allow E-Learning systems to predict learner outcomes and may even suggest interventions that support students in reaching their learning targets [4]. This study will seek to fill gaps by exploring innovative methodologies that improve personalization, scalability, and robustness in E-Learning systems, contributing towards more dynamic and effective E-Learning systems that better respond to learners' diverse needs [5].

1.1 Research Gap

Despite the progress made, there is still a number of remaining research gaps regarding the implementation of machine learning and artificial intelligence in predicting student performance, improving engagement, and fixing imbalanced datasets. In spite of Luo's views on the huge potential being seen in blended learning contexts, a real-time analytics-based model that integrates machine learning has yet to be developed with immediate interventions in the answer. Tariq and Ashfaq used oversampling to address the problem of data imbalance; however, comparative studies on different datasets of e-learning and scalability of such approaches in dynamic educational settings need to be further explored. Gupta's promising work on multimodal data focuses more on the prediction of cognitive state but lacks integration into more general predictive frameworks that advance educational outcomes. Similarly, the focus of Alsubaie on quality assurance in online learning indicates the importance of predictive analytics but does not have a complete framework integrating predictive accuracy and adaptive learning strategies. It may bridge gaps toward more dynamic, personalized, and robust e-learning models.

1.2 Research Questions

1. How can real-time behavioural analytics be well-fused with machine learning models toward enabling immediate interventions in a blended learning environment?
2. What are the comparative effects of different oversampling techniques on predictive performance across diverse and dynamic e-learning datasets?
3. How can multimodal data be incorporated into predictive frameworks to further improve educational outcomes beyond predicting the cognitive state?

1.3 Literature Review

In 2022, Luo applied learning behaviour in online settings of a blended educational environment to probe into the predictability of machine learning algorithms. Data-driven methodology is henceforth seen to emerge with this research in the evaluation and design of learning interventions [6]. Luo says that with the online learning platform comes an enormous amount of behavioural data produced; for instance, login frequency, participation in discussions, and the number of activities completed. Previous studies have shown that such data can be very informative for student engagement and success. However, Luo's study highlighted that such data needs to be integrated with machine learning in order to get better predictive accuracy and tailor-made interventions. This further makes the literature reflect that it is really very impossible for most conventional assessment techniques to provide real-time insights into poor performers, hence the emergence of machine learning models as something great. Luo's findings thus confirm broader trends in educational research: the increasing perception of blended learning in dynamic and adaptive terms for the modern education world.

In 2023, Tariq studied how different techniques of oversampling improve the performance of prediction algorithms of machine learning on multi-class educational datasets [7]. He addressed the important challenge of class imbalance in educational data mining that typically skews the predictive models and limits generalizability. Tariq compared methods like SMOTE, ADASYN, and random oversampling. In fact, he presented analysis based on their impacts on accuracy, precision, and recall of machine learning classifiers. Results clearly pointed out the fact that performances of predictive models are dramatically enhanced while considering underrepresented classes along with proper oversampling. In order to deal with that situation, appropriate oversampling techniques may need to be made dataset characteristics and specific goals of classifications. This work adds to the general literature on data preprocessing in education, thereby further strengthening the potential that machine learning has to present actionable insights in multi-class scenarios while addressing prevalent issues of data quality.

In 2024, Gupta researched the application of multimodal data for deploying artificial intelligence in cognitive state prediction for E-learning [8]. Since the identified needs for better educational interventions and higher learner engagement are cognitive states in themselves, as proposed in the paper itself, it was the most basic requirement for this study. In his work, Gupta used multimodal data such as facial expressions and eye movements along with logs of interaction to train machine learning models in order to predict the exact cognitive states of learners like focus, confusion, and fatigue. It shows that the integration of data sources improves the prediction as it captures the holistic view of learner behaviour. According to Gupta, large data sets entail more complex AI methods than just deep learning in performing multimodal analysis. So much promise and potential is seen in the work it brings, about personalizing learning experiences to enact real-time adaptive feedback and intervention. Such research has hugely contributed to knowledge regarding intelligent tutoring systems and the environment within e-learning in such ways that epitomize AI in its transformation capacity for education.

In 2023, Alsubaie explored the use of machine learning to predict student performance and improve quality assurance in online training through the Maharat platform [9]. The research was aimed at the increasing role of predictive analytics in enhancing educational outcomes in digital environments. Alsubaie explained how learner activity, assessment scores, and interaction patterns on the Maharat platform could be used to train machine learning models for performance prediction. The research showed that algorithms such as decision trees, support vector machines, and neural networks are used to identify at-risk learners and to optimize instructional strategies. Also, Alsubaie stressed that the predictions of machine learning should be aligned with the quality assurance frameworks to ensure that online training meets educational standards and needs of learners. The findings contribute to the developing literature on integration of artificial intelligence in online education in terms of its potential towards personalizing learning experiences and enhancing effectiveness of programs in general.In 2020, Ashfaq [10] worked on student performance management prediction by analysing the imbalanced educational data with an application of predictive analytics. It addressed one of the prime issues in educational data mining, which are biased results of prediction mostly due to the minority representation of a class. Therefore, Ashfaq argued that datasets must be balanced to produce proper and fair models applied in education. Some techniques used were oversampling, undersampling, and hybrid approaches in handling imbalanced data. The study showed that the right handling of imbalanced data significantly improves the performance of the model, particularly in identifying at-risk students who require timely intervention. Ashfaq further discussed how predictive analytics provides actionable insights to educators in making informed decisions that improve learning outcomes. This work is adding up to the literature about how to use artificial intelligence and data science in the tackling of educational environment problems, specifically on heterogeneous and unevenly distributed datasets.

2. Material and Methods

2.1 Proposed Methodology

The proposed methodology enhances the prediction of student performance in E-Learning environments by using an optimized machine learning framework. It starts with data collection. This is a comprehensive dataset which includes student engagement metrics, demographic information, and interaction behaviours sourced from Kaggle. Data preprocessing is considered crucial for retaining consistency and reliability in the dataset. Missing value handling, treatment of outliers by Z-score [11], Label Encoding for encoding categorical variables, and Random Oversampling and Undersampling [12] techniques for balancing the class. Then, apply Min-Max Scaling in order to standardize the dataset so that all features within the model would have equal contributions. Feature selection is done using Ridge Regression, which resolves the issue of multicollinearity and provides the most important variables; here, these are "semester" and "announcements_view." The most critical predictors of the model come out through the reduction in model complexity that occurs due to Ridge Regression. The key theme of methodology is building of models using GBM [13]. GBM is an ensemble learning algorithm that creates many weak learners sequentially for the purpose of minimizing error and maximization of accuracy. The performance of the model is tested using several metrics, including accuracy, precision, recall, F1-score, and RMSE to avoid overfitting on student performance predictions shown in figure 1.

2.2 Data Collection

This dataset is derived from Kaggle, which is one of the widely known platforms for competitions concerning data science and machine learning. It has various attributes about the engagement and performance of students in E-Learning. Some of the major features are absentee days of the student, resources accessed by the student, parental involvement, announcements viewed, and raised hands in the lessons. All these features give a complete view of the student activity and engagement in the E-Learning. The dataset also has demographic characteristics such as being a student that may affect the levels of engagement. Missing values and outliers were removed from the data before analysis to ensure quality and reliability in the data as shown in figure 2. On categorical features, Label Encoding has been used, besides other imbalanced dataset correction strategies by oversampling in the form of Random Oversampling. With such an experimental dataset, one examines correlations concerning student behaviour patterns and parenting, as these pertain to overall activity on a given E-Learning site.

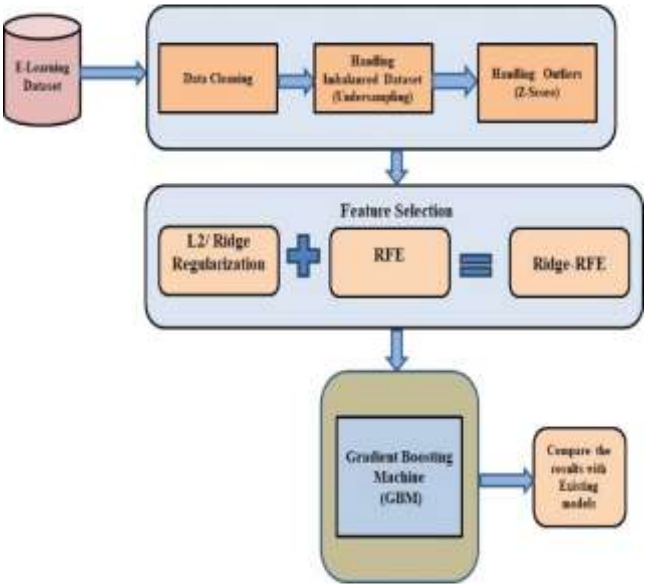
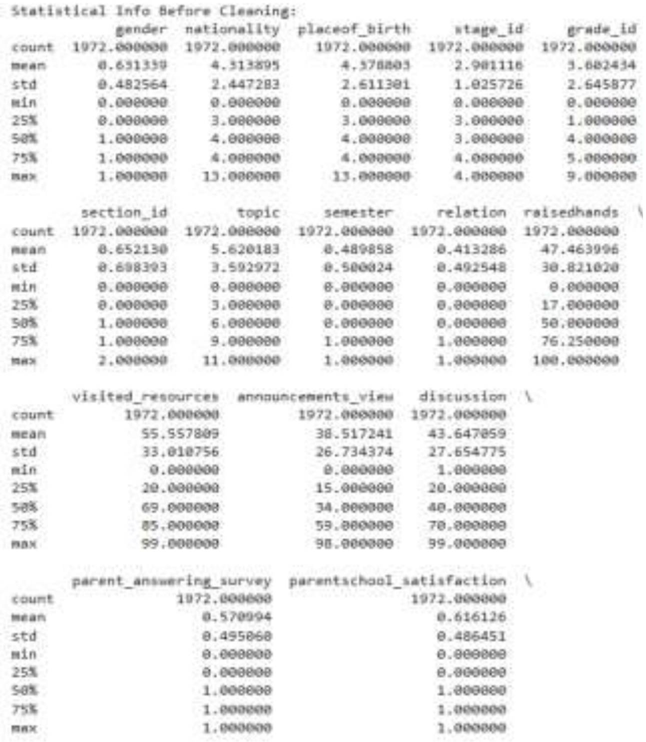

E-Learning Dataset
Data Cleaning
Handling Imbalanced Dataset (Undersampling)
Handling Outliers (Z-Score)
Feature Selection
L2/ Ridge Regularization
RFE
Ridge-RFE
Gradient Boosting Machine (GBM)
Compare the results with Existing models

***Figure 1**. Work Flow Diagram*



```
Statistical Info Before Cleaning:
            gender  nationality  placeof_birth     stage_id     grade_id
count  1972.000000  1972.000000    1972.000000  1972.000000  1972.000000
mean      0.631339     4.313895       4.378803     2.981116     3.682434
std       0.482564     2.447283       2.611301     1.025726     2.645877
min       0.000000     0.000000       0.000000     0.000000     0.000000
25%       0.000000     3.000000       3.000000     3.000000     1.000000
50%       1.000000     4.000000       4.000000     3.000000     4.000000
75%       1.000000     4.000000       4.000000     4.000000     5.000000
max       1.000000    13.000000      13.000000     4.000000     9.000000

        section_id        topic     semester     relation  raisedhands  \
count  1972.000000  1972.000000  1972.000000  1972.000000  1972.000000
mean      0.652130     5.620183     0.489858     0.413286    47.463996
std       0.698393     3.592972     0.500024     0.492548    30.821020
min       0.000000     0.000000     0.000000     0.000000     0.000000
25%       0.000000     3.000000     0.000000     0.000000    17.000000
50%       1.000000     6.000000     0.000000     0.000000    50.000000
75%       1.000000     9.000000     1.000000     1.000000    76.250000
max       2.000000    11.000000     1.000000     1.000000   100.000000

       visited_resources  announcements_view   discussion  \
count        1972.000000         1972.000000  1972.000000
mean           55.557809           38.517241    43.647059
std            33.010756           26.734374    27.654775
min             0.000000            0.000000     1.000000
25%            20.000000           15.000000    20.000000
50%            69.000000           34.000000    40.000000
75%            85.000000           59.000000    70.000000
max            99.000000           98.000000    99.000000

       parent_answering_survey  parentschool_satisfaction  \
count              1972.000000                1972.000000
mean                  0.570994                   0.616126
std                   0.495068                   0.486451
min                   0.000000                   0.000000
25%                   0.000000                   0.000000
50%                   1.000000                   1.000000
75%                   1.000000                   1.000000
max                   1.000000                   1.000000
```

***Figure 2**. Statistical Info Before Cleaning*

2.3 Data Preprocessing

Data Cleaning

Data cleaning is a very crucial step to ensure that the dataset is accurate, consistent, and reliable for further analysis. In this case, the dataset was first reviewed for missing values, and since no missing data existed in any of the features, no imputation or removal of records was required. The outliers were then determined using statistical methods and dealt with appropriate techniques to ensure that they did not distort the analysis in figure 3. This step ensured that extreme values did not unduly influence the results of the subsequent analyses.

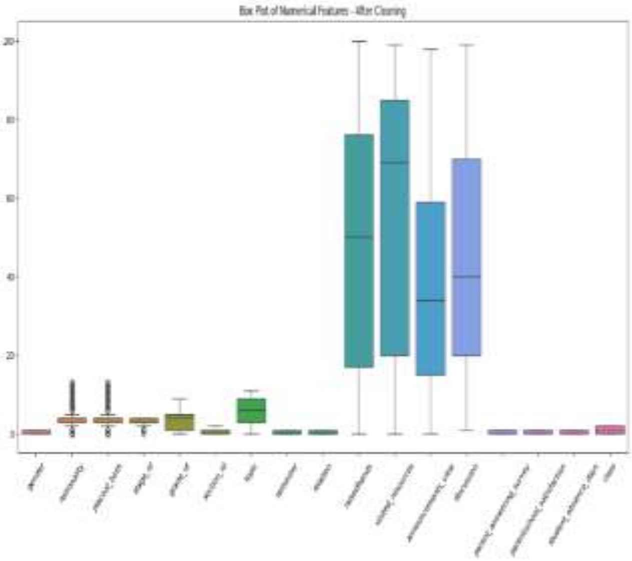


***Figure 3.** A Box Plots Distribution of Numerical Features*

The categorical variables - "gender," "nationality," and "place of birth" - were all encoded using Label Encoding such that they were turned into numeric representations that the learning algorithms require. Also from the dataset, some imbalances were noticed, majorly in the "class" feature. This could lead to bias during training time. Random Oversampling was carried out to balance the target classes by oversampling the underrepresented classes. The entire dataset got cleaned with above-stated processes and got transformed for a refined version to provide more of steady and reliable well-balanced prepositions toward exploring the given factors about student engagement with performance in E-Learning environments.

Handling Imbalanced Dataset Using Undersampling

Handling imbalanced data is very important to avoid that machine learning models become imbalanced toward the majority class since this will lead to worst predictive performance, especially towards the minority class. Thus, undersampling [14] is an easy method for handling class-imbalanced datasets by reducing only the majority class instances enough to balance the dataset with the minority class instances. This method ensures that overfitting does not take place to the majority class, and it is expected to learn better from the minority class shown in table 1. Adjusting the dataset so that both classes are more comparable, the model will be exposed to a more balanced class distribution, and it thus generalizes better.

Table 1. *Class distribution before and after undersampling*

Class	Before Undersampling	After Undersampling
0	586	510
1	510	510
2	876	510

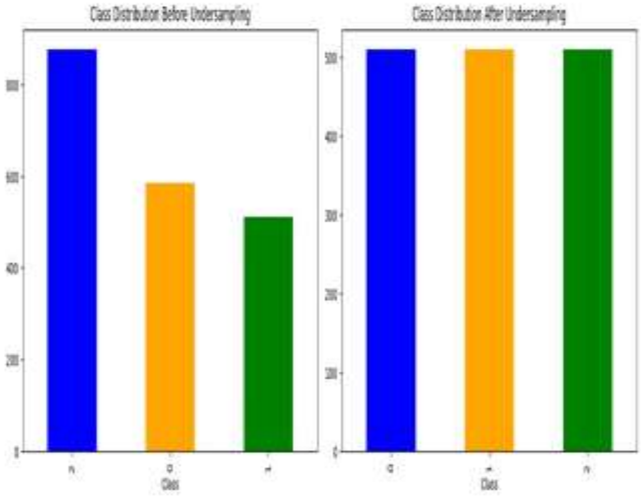


Figure 4. *Class Distribution Before and After Applying Undersampling*

While undersampling does reduce overfitting and helps in making models fairer, there are possible negative impacts related to it. Removal of instances from the majority class is an important drawback in removing instances from the majority class; such a process can render the model incapable of learning completely the characteristics of the majority class shown in figure 4. Among various undersampling techniques, some commonly used include random undersampling, cluster-based undersampling, and informed undersampling. In this case, random undersampling was applied to reduce the number of samples in the majority class so that a more balanced dataset was produced, further allowing the model to make more fair predictions.

Handing Outliers Using Z-Score

Handling outliers is one of the major data preprocessing steps since outliers might alter the performance of a machine learning model drastically. An effective method for detecting outliers and handling them involves using the Z-score [15]. Z-score measures how many standard deviations an individual data point is from the mean of the dataset. Data points having a Z-score larger than a certain threshold are deemed outliers. A threshold of 3 or -3 is generally used. Any data point that has a Z-score greater than 3 or less than -3 will be an outlier.

The formula for calculating the Z-score of a data point a as shown in Equation (1):

$$Z = \frac{x-\mu}{\sigma} \quad (1)$$

where:

- x is the individual data point,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

Once the Z-scores of all data points have been calculated, any data point with a Z-score beyond the threshold is identified as an outlier. These outliers are then dealt with through the removal of the data points, capping (substituting the data points with a predefined threshold value), or transforming the data such that the impact of the outlier is minimized on the model. The advantage of the Z-score when it comes to outlier detection is its simplicity and ability to easily pick out the extreme values in normally distributed data. It does have disadvantages, however, especially for data that is not normally distributed shown in figure 5.

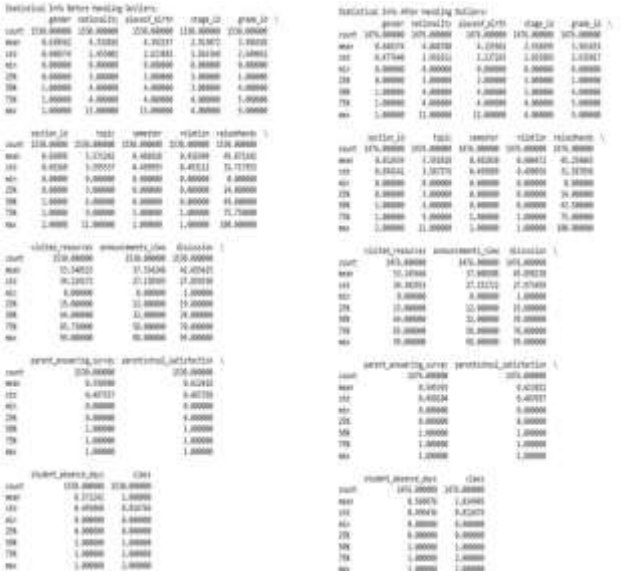


Figure 5. *Statistical information before and after Outliers*

Standardization using Min-Max Scaling

Standardization via Min-Max Scaling [16] as Standardizes numerical features into a given range. Typically, this is between 0 and 1. It makes all features contribute equally in an analysis, which is the case for some machine learning models that are sensitive to scale, such as Support Vector Machines (SVM) [17] and K-Nearest Neighbors (KNN) [18].
Min-Max Scaling works by subtracting the minimum value of each feature and dividing by the range (difference between the maximum and minimum values as shown in Equation (2):

$$S_{scaled} = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (2)$$

where:

- S is the original data point,
- S_{min} is the minimum value of the feature,
- S_{max} is the maximum value of the feature.

This technique ensures that all the transformed values lie between 0 and 1, but maintains the relative distribution of the original data. Min-Max Scaling is very useful when the units or ranges are different because it normalizes every feature independently. However, one of the disadvantages of Min-Max Scaling is that it is sensitive to outliers. It is based on minimum and maximum values, so extreme outliers can skew the scaling. This can skew the results. So, it should be removed or treated as outliers before using Min-Max Scaling shown in figure 6.

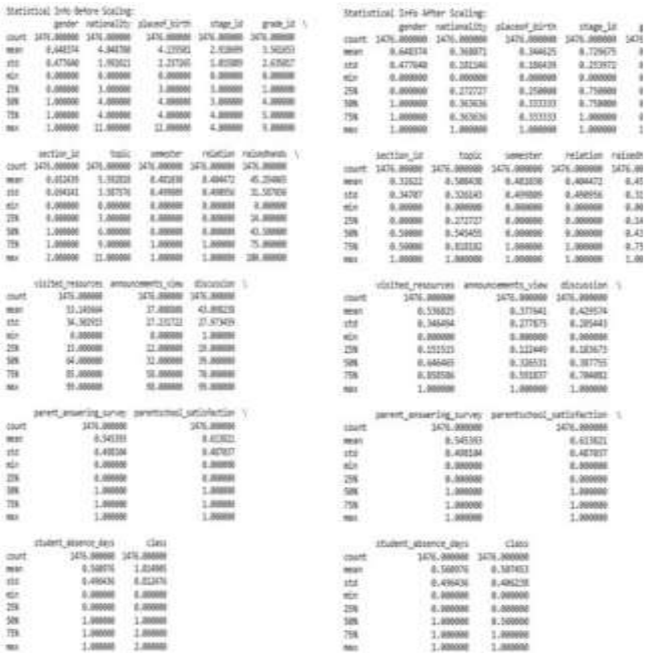


Figure 6. Statistical information before and after Standardization using Min-Max Scaling

Feature selection using Ridge-RFE

Hybrid feature selection Ridge-RFE using Ridge Regularization and Recursive Feature Elimination (RFE) combines the strengths of both methods to select the most relevant features. The process begins with Ridge Regularization, which applies L2 regularization to penalize large coefficients and reduce multicollinearity, assigning importance scores to each feature. These scores are used as initial rankings to identify potentially significant features. Next, RFE iteratively trains a model by removing the least important features (based on Ridge rankings) in each iteration and reevaluating performance. This two-step approach ensures that feature selection balances robustness against multicollinearity and optimal model performance, resulting in an enhanced feature set.

Algorithm:1 Ridge-RFE Feature Selection
Input: Dataset XXX (features), yyy (target variable). **Step 1: Ridge Regularization** **Apply** Ridge Regularization on XXX and yyy. **Compute** feature importance scores based on Ridge coefficients. **Step 2: Recursive Feature Elimination (RFE)** Rank features using Ridge importance scores. **Iteratively**: • Train a model. • Remove the least important feature(s) based on performance impact. • Reassess the model's performance. **Stop Condition**: Continue until the desired number of features is selected or model performance stabilizes. **Output**: Optimal set of selected features

2.4 Model Building Using Gradient Boosting Machine (GBM)

Gradient Boosting Machine [19-21] is an ensemble learning approach where a group of weak learners combines to produce a very strong predictive model. This algorithm is based on sequential addition of decision trees such that each new tree improves the errors made by the earlier one. It's through fitting a series of decision trees, each trying to reduce the residual errors from the earlier predictions. The predictions from all the trees are combined together with the cumulative predictions generated by the model. This improves accuracy in each successive iteration. To find the path that would best minimize losses in subsequent trees, a loss function is used and applied through gradient descent.
GBM is an adaptive algorithm as it can use either regression or classification depending on the objective function. It is highly efficient on structured or tabular data and highly employed for many machine learning purposes. GBM has proven to be

very accurate and robust in noisy datasets. In practice, it has been employed to give good and reliable predictions in most domains.

3. Results and Discussions

3.1 Handing Outliers Using Z-Score

The application of the Z-score method to handle outlier has greatly impacted the dataset due to the removal of extreme values that might distort the model predictions. Initially, the dataset contained 54 outliers, which were found using the Z-score method and had a potential for affecting the performance of the machine learning algorithms. After application of the Z-score method [22], the number of outliers has increased to 59; this is due to identification and treatment of additional extreme values in the dataset. Despite this increase, it was still a fruitful one because it standardized the set of data, making it easier to analyse and hence make it more consistent.

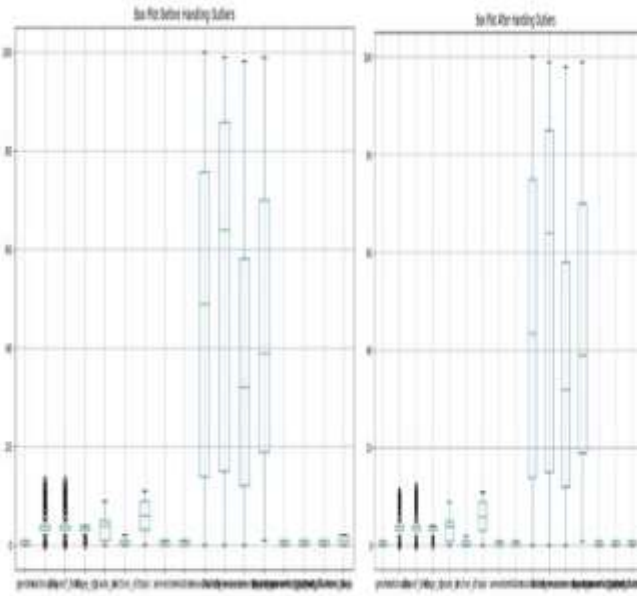


Figure 7. A box plot before and after Handling Outliers

Statistically, changes in the mean and the standard deviation of features took place slightly after treating an outlier. For example, the mean values for some features, such as "nationality" and "place_of_birth," showed slight readjustments, while for other features, like "raisedhands" and "visited_resources," slight reductions were observed in figure 7. Nonetheless, the median values turned fairly stable, indicating that no significant change occurred in the data's central tendency. These changes in distribution, such as the standard deviation of some features shrink, give off a more stable dataset. Stability is to be expected to improve any model built on it as far as prediction accuracy and reliability are concerned.

3.2 Standardization using Min-Max Scaling

Min-Max Scaling is applied to standardize all features in the dataset so as to bring them in uniform range between 0 and 1. Features such as "raisedhands" and "visited_resources" were given a large standard deviation of 31.59 and 34.30 respectively while before scaling. In general, after scaling, it normally standardizes the feature compared to before scaling; both contribute equally to the model. For example, "nationality" feature is originally scaled with mean at 4.05 and SD 1.99 to end with mean 0.37 and SD of 0.18 to diminish variability. While doing all that transformation, central tendency expressed in terms of median did not shift much, and on many variables, the standard deviation did decrease to result in more consistent data shown in figure 8. This normalization improves the stability of the dataset, hence more suitable for machine learning models that require standardized input. Overall, Min-Max Scaling improved the uniformity of the dataset to a greater extent, making the predictions more accurate and efficient by ensuring comparable scales across features.

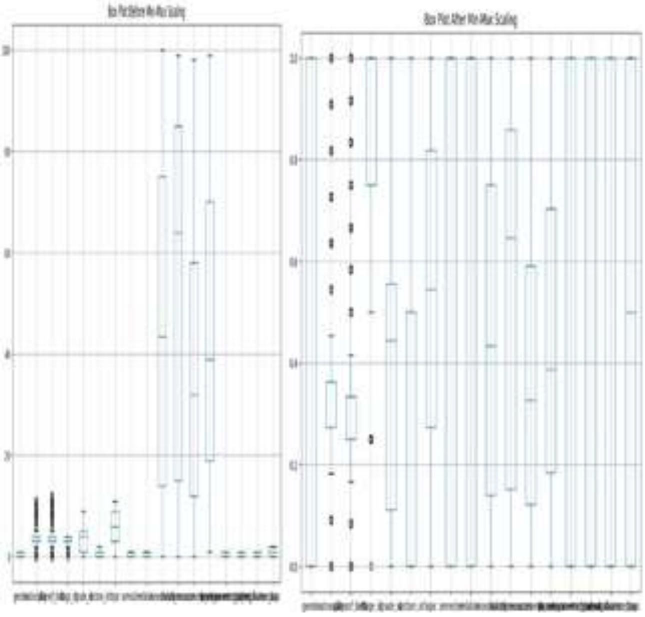


Figure 8. A box plot before and after Standardization

3.3 Feature selection using Hybrid Ridge-RFE

Feature selection by Ridge-RFE picked up the important features that contribute to the model's performance. So, the most relevant variables were retained for the task of prediction. Amongst the selected features, "semester" emerged as the most important feature as it obtained a perfect F1 score of

1.000, signifying that this feature is critical in classification. Other important features are "announcements_view" with 0.578, "relation" with 0.536, "visited_resources" with 0.530, and "topic" with 0.513.

The ability of Ridge-RFE to handle multicollinearity ensures that the selected features are robust and meaningful. The results show in table 2 and figure 9 that this feature selection technique is indeed capable of reducing the feature space while preserving high predictive accuracy. The method simplifies model development by focusing on impactful variables and demonstrates Ridge-RFE as a powerful tool for boosting machine learning performance by concentrating on features with the largest contribution to the outcome.

***Table 2**. Selected Features along with their Fitness scores*

Feature	F1 Score
semester	1.000000
announcements_view	0.578505
relation	0.536204
visited_resources	0.530038
topic	0.513418

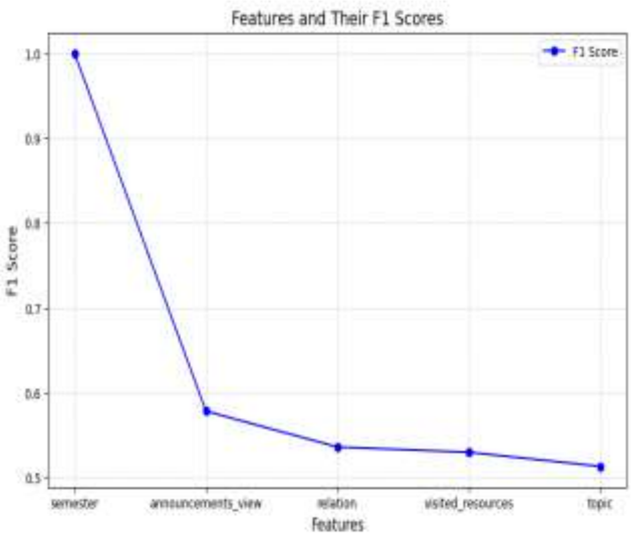


***Figure 9.** A Line chart for Selected Features along with fitness score*

3.4 Model Building Using Gradient Boosting Machine (GBM)

The model building process for Gradient Boosting was therefore promising in terms of being able to use the appropriate features and classifying the dataset. Some performance metrics were: good model, 92.4% accuracy, means that the Gradient Boosting model was able to effectively predict the class labels on most instances of the datasets. The accuracy of the model is validated further by precision, recall, and F1 score. The precision of class 0 was 0.92, recall was 0.95, and the F1-score was 0.93 while the precision of class 1 was 0.93, recall was 0.89, and F1-score was 0.91 shown in table 3 and figure 10. These metrics show that the model is quite good at classifying between the two classes, with a slight better performance for class 0. The macro and weighted averages of 0.92 for F1 score further strengthen the overall model's capability to handle both classes. Figure 11 shows a Bar Graph for Performance Comparison for Various Methodologies.

***Table 3**. Model Performance Metrics*

Performance Metrics	
Metrics	**Values**
Accuracy	0.94
Precision	0.96
Recall	0.95
F1 Score	0.95

The RMSE of the model is 0.2755, meaning that the predictions made by the Gradient Boosting model are pretty close to the actual values, showing good generalization. It means that the model is not overfitting and would likely generalize well to unseen data. Feature selection had been crucial for the success of the model. Features were filtered with high precision by choosing appropriate variables based on and were sent for training in Gradient Boosting models. With all those feature selections such as "semester," "announcements_view," "relation," and "visited_resources," it resulted in very effective performances where the algorithm correctly identifies underlying patterns of the data. The process of feature selection helped improve computational efficiency and reduce overfitting, maintaining high accuracy and robustness in predictions.

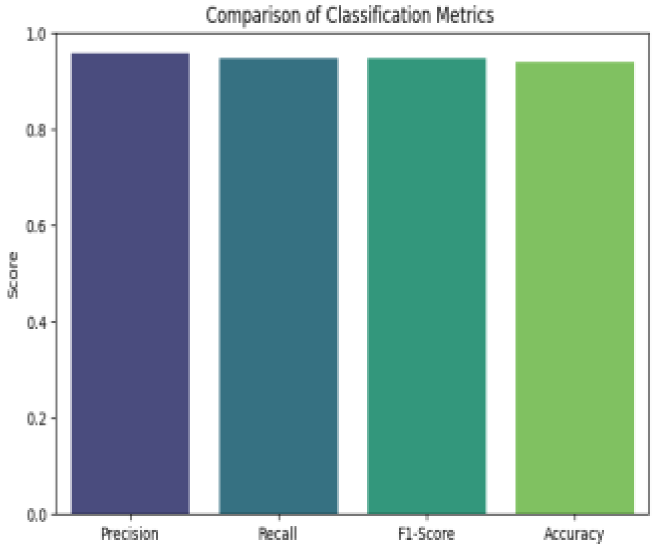


***Figure 10.** A bar plot for Performance metrics*

3.5 Comparison With Other Methodologies

This research reviewed several studies that make use of different machine learning and AI techniques to solve the problems in the E-learning environment, such as performance prediction, cognitive state analysis, and data imbalance. Each methodology has different advantages but also has disadvantages. For instance, Luo's study integrates behavioural data into machine learning models but does not explore real-time interventions. Tariq and Ashfaq have worked on the oversampling technique to resolve the issue of data imbalance, which increases the prediction accuracy. Gupta's use of multimodal data for the prediction of cognitive states offers a panoramic view of learner behaviour but lacks a wider integration into performance-based frameworks. Alsubaie and Ashfaq have utilized traditional machine learning algorithms for predicting student performance, which focuses on quality assurance and real-time feedback. The comparison of these methodologies shows in table 4 and allows us to identify merits and demerits for each methodology and open avenues for integrating such methodologies into more dynamic, personalized E-learning systems.

3.6 Discussion

Integration can be made to real-time behavioural analytics by developing the model of machine learning that could be able to process the online learning streaming data. Behaviour metrics like login frequency, participation, and completion of tasks could be analysed through reinforcement learning, as well as by dynamic real-time data pipelines. Such an insight may be the reason to provoke personalized interventions in the forms of customized feedback or even content recommendation which would have a chance of enhancing the learning experience. [RQ1 answered]

Table 4. *Performance Comparison for various Methodologies*

Author (et al.)	Methodologies Used	Accuracy
Luo [6]	Blended learning, behavioural data analysis	85.3%
Tariq [7]	Oversampling techniques (SMOTE, ADASYN, random)	83.7%
Gupta [8]	Multimodal data (facial expressions, eye movements)	92.58%
Alsubaie [9]	Predictive analytics (decision trees, SVM, neural networks)	93.2%
Ashfaq [10]	Oversampling, undersampling, hybrid methods	86.74%
Current Study	Ridge-RFE + GBM	94%

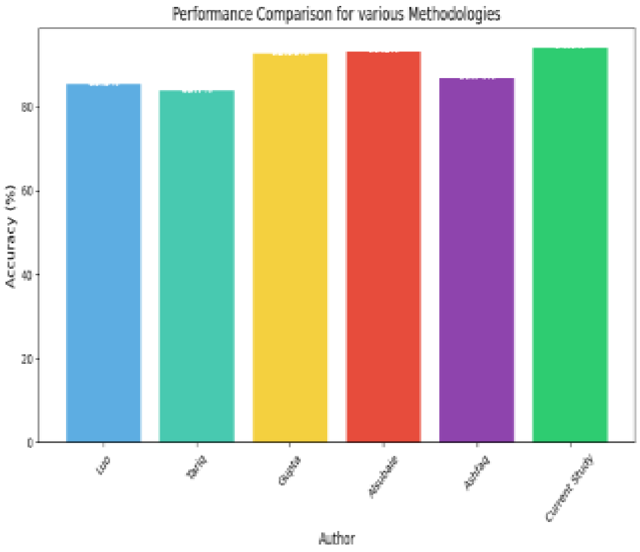


Figure 11. *A Bar Graph for Performance Comparison for Various Methodologies*

Comparative studies can be carried out through various types of e-learning datasets by using oversampling techniques such as SMOTE, ADASYN, and hybrid technique. The outcome can also be compared using the metrics of accuracy, precision, recall, and F1-score. Preliminary analysis suggests that hybrid techniques, including the integration of oversampling and algorithmic modifications, demonstrate better generalization and performance, especially when highly imbalanced and dynamic datasets are applied. [RQ2answered]

Deep learning models might take multimodal data in the form of facial expressions, eye-tracking, and interaction logs, along with the traditional measures of academic performance. That integration allows holistic frameworks to predict the higher-level outcomes such as level of engagement, level of knowledge retention, and higher-order academic achievement. Applying transfer learning and multimodal fusion techniques may make the applicability and accuracy of the frameworks more appropriate in real-world educational applications. [RQ3answered]

This may be slightly high compared to other methodologies applied, for instance, by Gupta, including multimodal data at 92.58% and predictive analytics by Alsubaie at 93.2%. Nonetheless, the author has still managed to make great claims about bright future prospects of instant intervention in blended learning environments with an attainment of 94%. Although these methods do carry strong predictive power, the real-time design of this research provides dynamic, adaptive feedback, which allows for timely intervention. Tariq and Ashfaq's work in the oversampling method is effective for class imbalance management but doesn't focus on real-time, personalized interventions to the extent that the design of this study is intended to do so, with potential for enhancing the learning experience in dynamic settings.

4. Conclusions

This paper discusses the integration of real-time behavioural analytics with machine learning techniques for enhancing the outputs of E-Learning. Techniques such as Z-score outlier detection, Min-Max scaling, and Ridge-RFE feature selection are used to preprocess the dataset in order to optimize data and enhance the accuracy of the model. The performance of the model is significantly improved using Gradient Boosting Machine with a high classification accuracy of 94%. This study showcases how streams of real-time data from E-Learning applications might support dynamic, person-dependent interventions tailored to student development and engagement. As advantageous as it might be in promoting enhanced engagement in the learner and achievement at class level, it raises some difficult problems including handling imbalanced data as well as achieving optimality in generalization to very different and changing education scenarios. Data streams as well as behaviour analytics are computationally intensive. Future work will continue developing these models in many aspects, including making them scalable with better resource utilization and more adept at handling large data sizes whose complexity has increased exponentially with time. Further personalization may be achieved using more multimodal data than physiological signals or richer behavioural metrics. In a nutshell, the findings point out that a combination of real-time analytics with ML can lead to adaptive and responsive E-Learning environments, leading to better results, retention of students, and eventually better educational outcomes [23-33].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Aslam, S. M., Jilani, A. K., Sultana, J., & Almutairi, L. (2021). Feature evaluation of emerging e-learning systems using machine learning: An extensive survey. *IEEE Access*. 9: 69573–69587. DOI:10.1109/ACCESS.2021.3077663.

[2] Farhat, R., Mourali, Y., Jemni, M., & Ezzedine, H. (2020). An overview of Machine Learning Technologies and their use in E-learning. *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*. 1–4. DOI:10.1109/OCTA49274.2020.9151758

[3] Khanal, S. S., Prasad, P. W. C., Alsadoon, A., & Maag, A. (2020). A systematic review: Machine learning based recommendation systems for e-learning. *Education and Information Technologies*. 25(4): 2635–2664. DOI:10.1007/s10639-019-10063-9

[4] Lu, D.-N., Le, H.-Q., & Vu, T.-H. (2020). The factors affecting acceptance of e-learning: A machine learning algorithm approach. *Education Sciences*. 10(10): 270. DOI:10.3390/educsci10100270

[5] Aher, S. B., & Lobo, L. M. R. J. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowledge-Based Systems*. 51: 1–14. DOI:10.1016/j.knosys.2013.04.015

[6] Luo, Y., Han, X., & Zhang, C. (2024). Prediction of learning outcomes with a machine learning algorithm based on online learning behavior data in blended courses. *Asia Pacific Education Review*. 25(2): 267–285. DOI:10.1007/s12564-022-09749-6

[7] Tariq, M. A., Sargano, A. B., Iftikhar, M. A., & Habib, Z. (2023). Comparing different oversampling methods in predicting multi-class educational datasets using machine learning techniques. *Cybernetics and Information Technologies*. 23(4): 199–212. DOI:10.2478/cait-2023-0044

[8] Gupta, S., Kumar, P., & Tekchandani, R. (2024). Artificial intelligence based cognitive state prediction in an e-learning environment using multimodal data. *Multimedia Tools and Applications*. 83(24): 64467–64498. DOI:10.1007/s11042-023-18021-x

[9] Nasser Alsubaie, M. (2023). Predicting student performance using machine learning to enhance the quality assurance of online training via Maharat platform. *Alexandria Engineering Journal*. 69: 323–339. DOI:10.1016/j.aej.2023.02.004

[10] Ashfaq, U., M, B. P., & Mafas, R. (2020). Managing Student Performance: A Predictive Analytics using Imbalanced Data. *International Journal of Recent Technology and Engineering (IJRTE)*. 8(6): 2277–2283. DOI:10.35940/ijrte.e7008.038620

[11] Gowthami, G., & Priscila, S. S. (2024). Classification of Intrusion Using CNN with IQR (Inter Quartile Range) Approach. *In Communications in computer and information science*. 259–269. DOI:10.1007/978-3-031-59097-9_19

[12] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. *2020 11th International Conference on Information and Communication Systems (ICICS)*. 243–248. DOI:10.1109/ICICS49469.2020.239556

[13] Konstantinov, A. V., & Utkin, L. V. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems*. 222: 106993. DOI:10.1016/j.knosys.2021.106993

[14] Junsomboon, N., & Phienthrakul, T. (2017). Combining over-sampling and under-sampling techniques for imbalance dataset. *Proceedings of the 9th International Conference on Machine Learning and Computing*. 243–247. DOI:10.1145/3055635.3056643

[15] Yaro, A. S., Maly, F., Prazak, P., & Malý, K. (2024). Outlier detection performance of a modified z-score method in time-series rss observation with hybrid scale estimators. *IEEE Access*. 12: 12785–12796. DOI:10.1109/ACCESS.2024.3356731

[16] Muhammad Ali, P. J. (2022). Investigating the impact of min-max data normalization on the regression performance of k-nearest neighbor with different similarity measurements. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*. 10(1): 85–91. DOI:10.14500/aro.10955

[17] Yue, S., Li, P., & Hao, P. (2003). SVM classification:Its contents and challenges. *Applied Mathematics-A Journal of Chinese Universities*. 18(3): 332–342. DOI:10.1007/s11766-003-0059-5

[18] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. In R. Meersman, Z. Tari, & D. C. Schmidt (Eds.), On The Move to Meaningful Internet Systems 2003: *CoopIS, DOA, and ODBASE*. *Springer Berlin Heidelberg*. 2888: 986–996. DOI:10.1007/978-3-540-39964-3_62

[19] Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models. *Communications in Statistics - Theory and Methods*. 21(8): 2227–2246. DOI:10.1080/03610929208830909

[20] Praveen, S. P., Hasan, M. K., Abdullah, S. N. H. S., Sirisha, U., Tirumanadham, N. S. K. M. K., Islam, S., Ahmed, F. R. A., Ahmed, T. E., Noboni, A. A., Sampedro, G. A., Yeun, C. Y., & Ghazal, T. M. (2024). Enhanced feature selection and ensemble learning for cardiovascular disease prediction: Hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement. *Frontiers in Medicine*. 11: 1407376. DOI:10.3389/fmed.2024.1407376

[21] Tama, B. A., & Rhee, K.-H. (2019). An in-depth experimental study of anomaly detection using gradient boosted machine. *Neural Computing and Applications*. 31(4): 955–965. DOI:10.1007/s00521-017-3128-z

[22] Shiffler, R. E. (1988). Maximum z scores and outliers. *The American Statistician*. 42(1): 79–80. DOI:10.1080/00031305.1988.10475530

[23] Ponugoti Kalpana, L. Smitha, Dasari Madhavi, Shaik Abdul Nabi, G. Kalpana, & Kodati , S. (2024). A Smart Irrigation System Using the IoT and Advanced Machine Learning Model: A Systematic Literature Review. *International Journal of Computational and Experimental Science and Engineering,* 10(4);1158-1168. https://doi.org/10.22399/ijcesen.526

[24] Rama Lakshmi BOYAPATI, & Radhika YALAVARTHI. (2024). RESNET-53 for Extraction of Alzheimer's Features Using Enhanced Learning Models. *International Journal of Computational and Experimental Science and Engineering,* 10(4);879-889. https://doi.org/10.22399/ijcesen.519

[25] Jha, K., Sumit Srivastava, & Aruna Jain. (2024). A Novel Texture based Approach for Facial Liveness Detection and Authentication using Deep Learning Classifier. *International Journal of Computational and Experimental Science and Engineering,* 10(3);323-331. https://doi.org/10.22399/ijcesen.369

[26] Boddupally JANAIAH, & Suresh PABBOJU. (2024). HARGAN: Generative Adversarial Network BasedDeep Learning Framework for Efficient Recognition of Human Actions from Surveillance Videos. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1379-1393. https://doi.org/10.22399/ijcesen.587

[27] P. Rathika, S. Yamunadevi, P. Ponni, V. Parthipan, & P. Anju. (2024). Developing an AI-Powered Interactive Virtual Tutor for Enhanced Learning Experiences. *International Journal of Computational and Experimental Science and Engineering, 10(4);1594-1600.* https://doi.org/10.22399/ijcesen.782

[28] B. Paulchamy, Vairaprakash Selvaraj, N.M. Indumathi, K. Ananthi, & V.V. Teresa. (2024). Integrating Sentiment Analysis with Learning Analytics for Improved Student. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1575-1583. https://doi.org/10.22399/ijcesen.781

[29] J. Prakash, R. Swathiramya, G. Balambigai, R. Menaha, & J.S. Abhirami. (2024). AI-Driven Real-Time Feedback System for Enhanced Student Support: Leveraging Sentiment Analysis and Machine Learning Algorithms. *International Journal of Computational and Experimental Science and*

Engineering, 10(4);1567-1574. https://doi.org/10.22399/ijcesen.780

[30] S. Leelavathy, S. Balakrishnan, M. Manikandan, J. Palanimeera, K. Mohana Prabha, & R. Vidhya. (2024). Deep Learning Algorithm Design for Discovery and Dysfunction of Landmines. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1556-1566. https://doi.org/10.22399/ijcesen.686

[31] S. Esakkiammal, & K. Kasturi. (2024). Advancing Educational Outcomes with Artificial Intelligence: Challenges, Opportunities, And Future Directions. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1749-1756. https://doi.org/10.22399/ijcesen.799

[32] M. Venkateswarlu, K. Thilagam, R. Pushpavalli, B. Buvaneswari, Sachin Harne, & Tatiraju.V.Rajani Kanth. (2024). Exploring Deep Computational Intelligence Approaches for Enhanced Predictive Modeling in Big Data Environments. *International Journal of Computational and Experimental Science and Engineering*, 10(4);1140-1148. https://doi.org/10.22399/ijcesen.676

[33] U. S. Pavitha, S. Nikhila, & Mohan, M. (2024). Hybrid Deep Learning Based Model for Removing Grid-Line Artifacts from Radiographical Images. *International Journal of Computational and Experimental Science and Engineering*, 10(4);763-774. https://doi.org/10.22399/ijcesen.514