ISSN 2278-3091
Volume 12, No.5, September - October 2023
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse041252023.pdf
https://doi.org/10.30534/ijatcse/2023/041252023

Predictive Modeling for Enhancing Academic Performance in Nigerian Polytechnic Education

Unyime Edet[1], Sunday Obot Iwok[2]
[1]Department of Computer Science, Akwa Ibom State Polytechnic, Ikot Osurua, Nigeria, unyime.edet@akwaibompoly.edu.ng
[2]Department of Computer Science, Akwa Ibom State Polytechnic, Ikot Osurua, Nigeria, sundayiwok@yahoo.com

Received Date : September 01, 2023 Accepted Date: September 28, 2023 Published Date: October 06, 2023

ABSTRACT

This study presents a machine learning-based approach to enhance the classification and optimization of students' academic performance within Nigeria's polytechnic education system. The polytechnic system is pivotal in providing technical and vocational education, but challenges persist in nurturing students' academic achievement. This article explores the complexities influencing academic performance and proposes strategies for improvement using machine learning algorithms. The research utilizes linear and support vector regression models to predict students' cumulative grade point averages (CGPA). A dataset from Akwa Ibom State Polytechnic, Ikot Osurua, comprising total courses, credit units, department, and previous grade point average (GPA), is employed for model development and evaluation. Both models achieve similar predictive performance, but linear regression slightly outperforms support vector regression. The results highlight the significant role of variables like total courses, the type of academic department, and previous GPA in predicting CGPA. This study offers a valuable tool for assessing and improving students' academic performance in Nigeria's polytechnic education system, with potential for broader applications in higher education. Further research involves expanding the dataset and considering additional factors beyond result records to enhance the model's robustness and applicability.

Keywords: Predictive modelling, academic performance enhancement, Nigerian polytechnic education.

1. INTRODUCTION

This study is a machine learning-based system aimed at adopting a model that will facilitate the classification and optimization of students' academic performance in the polytechnic education system in Nigeria.

The process through which a person gets or transmits knowledge, information, skills, experiences, talents, and attitudes essential for an active and meaningful existence in society is called education [1]. Education is the cornerstone of development and progress, serving as a pathway to personal growth and societal advancement. In Nigeria, the polytechnic educational system is crucial in providing technical and vocational education to a diverse pool of students. However, a critical concern remains: How can students' academic performance be effectively nurtured and elevated within this unique educational framework? This article explores the multifaceted dynamics that influence students' academic performance in Nigeria's polytechnic system and suggests strategies for improvement using machine learning-based algorithms. The widespread consensus is that education is the key that opens the door to any society's social, economic, political, and technical growth. However, a lot depends on the substance and quality of it. The education system in Nigeria is a diverse and complex framework that reflects the country's cultural, historical, and socioeconomic dynamics. It is structured into several levels, each with unique characteristics and challenges. Tertiary education is one of such levels and encompasses three categories: universities, polytechnics, and colleges of education and monotechniques. Our focus of this paper is to explore strategies by which students' performance can be improved in the polytechnic education system in Nigeria. Polytechnic education is also known as technical education, which is defined as teaching a practically oriented skill or process at a level between that of a skilled craftsperson and a professional scientist or engineer [2]. Therefore, polytechnics offer technical and vocational education, often focusing on practical skills and workforce readiness.

New students frequently need help adapting to the rigorous academic standards when enrolling in high-level educational institutions. 20 to 30 percent of students drop out of postsecondary institutions after their first two years. The stress and challenges of transitioning to university or polytechnic life are suggested as a possible cause of this high dropout rate [3]. The academic sector is one of the main areas where new students face difficulties. Some of the factors that cause students to drop out of school include their inability to balance their social lives with their academic lives as a result of their sudden independence from their parents, difficulties attending classes, maintaining their grade level, meeting lecturers outside of class, adjusting to a

new learning environment, and confusion throughout study they should enroll in [4], [5], [6], and [7]. Therefore, those working in higher education must be aware of the difficulties first-year students experience. It is critical to comprehend these challenges since student achievement generally emphasizes the necessity of transition during the first year of college. Advisors may make wise judgments to help students have an excellent dynamic transition to college by using the right resources, such as the ability to estimate students' cumulative grade point average (CGPA) in their first year. The student and parents can choose the curriculum of their choice with more knowledge if it is possible to predict the Student's CGPA in the first academic year.

The global benchmark for measuring academic success has been the cumulative grade point average (CGPA). The influence of various departmental academic requirements and how they affect a student's academic performance are highlighted by this model through the prediction of a student's CGPA. This study uses previous results statistics to explore two machine learning techniques for predicting students' academic performance.

Reference [8] discussed data gathering methods for educational data mining in Nigeria in a paper on 'Data Collection Experience on Educational Data Mining in Nigeria'. The researcher underlined that learning analytics and educational data mining support techniques for identifying students' learning patterns by examining the many data types accessible in educational settings. In the paper, the author detailed the challenges of gathering data at two public institutions in Nigeria's Bayelsa State. The researcher suggested that to promote the benefits of data mining in education, scholars should be sufficiently educated and encouraged on the proper method of data collection and storage accessibility.

Reference [9] explores continuously applying EDM techniques to improve student academic performance at a higher learning level. In their research, they examined a predictive system using a machine learning framework that can be adopted for early prediction of Student academic performance. Identifying weak performance and offering the required rehabilitation to prevent school dropouts and support top achievers is critical. Their article examines specific characteristics of the 103 first-year Computer Science majors at the University of Nigeria, Nsukka. Their research used the Feature selection technique to filter feature variables from many independent variables affecting student performance. A Decision Tree machine learning algorithm was utilized for both training and testing. It was discovered that the datasets used to train the model impacted accuracy. On the same method, two distinct datasets achieve varying degrees of accuracy.

Five machine learning techniques, namely J48, Random Forest, Multilayer perceptron, decision tree and IB1, were used in the analysis [10]. In their work, they investigated the effectiveness of machine learning algorithms in determining the effects of parental education, gender, economics, and region on students' academic performance. The work was modelled as a classification problem, and the performance evaluation of such models used in their research was based on standard classification assessment model techniques such as confusion matrix, accuracy level and execution time.

Subsequently, small and midsized academic institutions, especially those for which most of their programmes constitute graduate and post-graduate courses, have small datasets of student records for analysis. Assessing student performance using machine learning techniques with a limited dataset was examined [11]. This study also investigated the potential to recognize the primary markers within the small dataset, which were employed in formulating the predictive model using visualization and clustering methods. The finest indicators were inputted into numerous machine learning algorithms to evaluate their suitability for constructing the most precise model. From the selected algorithms, the outcomes illustrated the clustering technique's proficiency in pinpointing key markers within limited datasets. The principal discoveries of this investigation showcased the efficiency of support vector machines and linear discriminant analysis algorithms in training with small dataset dimensions and attaining satisfactory classification precision and reliability test percentages.

A study was carried out on the modelling, predicting, and categorising student academic performance using artificial neural networks [12]. The study described a methodology that combined traditional statistical analysis with neural network modelling and student performance prediction. The variables that most likely influence the students' performance are found using conventional statistical computations. Eleven input variables, two levels of hidden neurons, and one output layer made up the neural network model. The backpropagation training rule was implemented using the Levenberg-Marquardt algorithm. The area under the receiver operating characteristics (ROC) curve, the confusion matrix, and the error histogram were used to assess the neural network model's performance with an 84.8% accuracy.

2. METHODOLOGY

The methods and approaches used in this research are outlined in the following section.

2.1 Data Mining Research

Numerous engineering, science, business, medicine, and education applications focus on data mining research, which has attracted much attention [13]. Data mining basic research is developing due to society's growing interest in knowledge discovery [14]. In the last ten years, the amount of digital data accessible to humans has multiplied, but the tools, scientists, and engineers needed to evaluate this growing volume of data have remained unchanged [15]. Data mining research aims to formulate, analyze, and extract knowledge and information from large amounts of

unstructured data. Data mining research studies, tools and procedures that may be applied to find patterns, changes, relationships, and structures with statistically significant data.

2.2 Research Framework

The proposed model framework is shown in Figure 1. The general concept of the proposed system is to develop and present a hyper-machine learning algorithm for predicting students' performance in terms of their cumulative grade points average at the end of their first academic year based on departmental academic requirements.

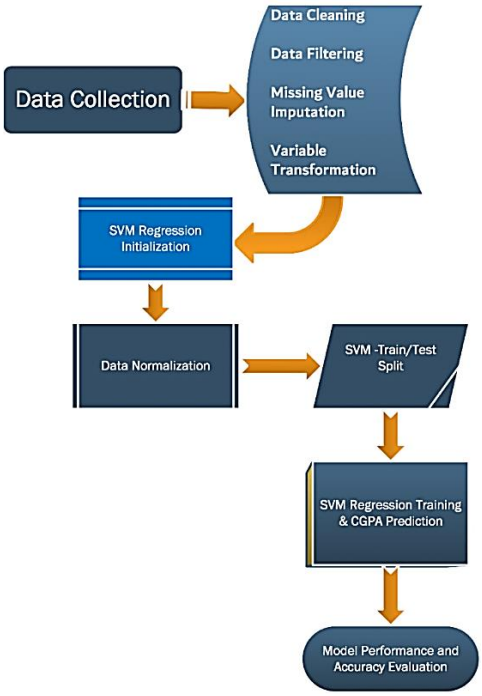

Figure 1: Proposed System Framework

A description of some of the components of the proposed system framework is presented as follows:

2.2.1 Data Collection

Data collection is a systematic procedure used in research to obtain, measure, or record information or observations. The data used in this research was obtained from the examination repository of the Akwa Ibom State Polytechnic, Ikot Osurua, Ikot Ekpene, Akwa Ibom State, Nigeria. The dataset comprises the second-semester results for the first-year students of the National Diploma programme in four departments collated over three years in the science faculty. The key reason for using data from multiple departments was to enhance model validity and reliability. Model reliability measures the global accuracy of the model performance, while model validity ensures that a developed model is significant across various data points. Contrary to the issue of data validity associated with secondary data sources, the nature of data used in this research, the student's semester results are certified by the institution, thereby having a well-established high degree of reliability and validity and not needing to be reexamined by the researcher. Furthermore, using the result dataset informed the adopted system design approach and provided a baseline for which the proposed system is designed and implemented.

2.2.2 Data Preprocessing

Data preprocessing is crucial in machine learning projects since it impacts the model's performance, validity, and ability to provide accurate results. This is supported by machine learning relying on provided data to make decisions or anticipate outcomes. The process of transforming, converting, or structuring raw data into a format suitable for developing machine learning models and statistical analysis is known as data preprocessing. In this research, the following data preprocessing methods were applied to the raw dataset:

1. Data Cleaning: This approach recognizes and manages data observations with missing or null values. This process is additionally known as missing value imputation. This approach replaces missing or invalid data observations with the mean of the feature variable or removes them entirely from the dataset.
2. Data Transformation: This method converted feature variables into acceptable forms for machine learning applications. Although there are many ways to change data, this study deployed variable transformation, normalization, and standardization.
3. Data Filtering: The technique of data filtering is employed in feature extraction. When extracted information from the raw dataset is not relevant to the study; as a result, feature variables and crucial and required observations were retrieved through the use of data filtering.

2.2.3 Project Dataset

The project dataset used in this study is a multivariate with 649 observations with eight features. The dataset was collected over three years for the second-semester academic calendar. The features and description of the dataset are presented in Table 1.

Table 1: Features of result dataset and its categories

S/N	Attributes	Description	Data Type
1	Total number of courses	Total number of courses offered in the current semester	Integer
2	Total number of credit unit	Total number of credit units in the current semester	Integer
3	Total number of three credit unit courses	Total number of courses with three credit units	Integer
4	Total number of four credit unit courses	Total number of courses with four credit units	Integer

5	Department	Student's department (Electrical & Electronics Engineering, Computer Science, Science Laboratory Technology and Statistics)	Categorical
5	Previous GPA	Previous Grade Point Average from Semester One	Float
6	CGPA	Cumulative Grade Point Average in the current semester	Float

2.2.4 Model Prediction

Machine learning algorithms can predict continuous data correctly for new parameters based on historic training datasets. This study considered two machine learning algorithms for training and prediction: linear regression and support vector machine. Since this is a regression problem, the CGPA attribute was chosen as the dependent variable. The two regression techniques were trained to predict the Student's CGPA. 80% of the data were used for the training, while 20% for testing. A random state of 0 was used for each model to obtain the same set of train-test data splits. A summary of the dataset of all the variables excluding GNS 102, GNS 121 and Department are shown in Figure 2. This summary describes the central tendency of the data based on the mean, median, count, mode, and standard deviation.

	T_no_courses	T_CU	T_3_CU	T_4_CU	P_GPA	CGPA
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	8.246533	23.493066	4.916795	1.329738	2.384472	2.268886
std	0.431325	0.862650	1.248151	0.845683	0.456249	0.410986
min	8.000000	23.000000	4.000000	0.000000	1.021739	0.717391
25%	8.000000	23.000000	4.000000	1.000000	2.097826	2.027174
50%	8.000000	23.000000	4.000000	2.000000	2.326087	2.190217
75%	8.000000	23.000000	5.000000	2.000000	2.833333	2.603261
max	9.000000	25.000000	7.000000	2.000000	3.826000	3.385870

Figure 2: Descriptive statistics of project dataset

2.2.5 Linear Regression

To predict a student's CGPA, a linear regression method was used. This method established a relationship between the predictor variables and the dependent variable. The dependent variable is the CGPA, while the independent variables are the features or predictors in the dataset. The intercept of the regression line is the predicted value when all the predictors' value is zero. The fitted values are the estimated output derived from the regression line, whereas the regression coefficient is also referred to as the slope of the regression line. The discrepancy between the observed and fitted values is residual or error. The linear regression is defined as:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + e\#(1)$$

The idea of fitted values is crucial to regression analysis. Since data seldom falls perfectly on a line, an explicit error component, e_i, must be included in the regression equation. The following are the CGPA predictions:

$$\hat{Y} = \widehat{b_0} + \widehat{b_1}X_1 + \widehat{b_2}X_2 + \widehat{b_3}X_3 + \widehat{b_4}X_4 + \widehat{b_5}X_5 + \widehat{b_6}X_6 + \hat{e}\#(2)$$

The notation $b_0 \rightarrow b_6$ indicates the estimated coefficients of the predictors. The error is computed as:

$$\hat{e}_i = Y_i - \hat{Y}_i\#(3)$$

The ordinary least squares (OLS) approach fits the data to the model. OLS is a technique for reducing the total squared residuals. The regression line, also known as the residual sum of squares (RSS), is the estimate that minimizes the sum of squared residual values and is expressed in the following equation:

$$RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n}\left(Y_i - \widehat{b_0} + \widehat{b_1}X_1 + \widehat{b_2}X_2 + \widehat{b_3}X_3 + \widehat{b_4}X_4 + \widehat{b_5}X_5 + \widehat{b_6}X_6\right)^2 \ \#(4)$$

The estimated $b_0 \rightarrow b_6$ are the values that minimize RSS.

2.2.6 Support Vector Regression (SVR)

SVR is another technique used for model prediction and will be derived from a geometrical perspective, using the one-dimensional data of previous GPA vs. CGPA presented in Figure 3.

To discover the narrowest tube centred on the surface while reducing the prediction error or the gap between the predicted and desired outputs, SVR formulates this function approximation issue as an optimization problem. The first condition produces the objective function in Equation 5, where $||w||$ is the size of the average vector to the surface being approximated.

$$\min_{w} \frac{1}{2} \| w \|^2 \#(5)$$

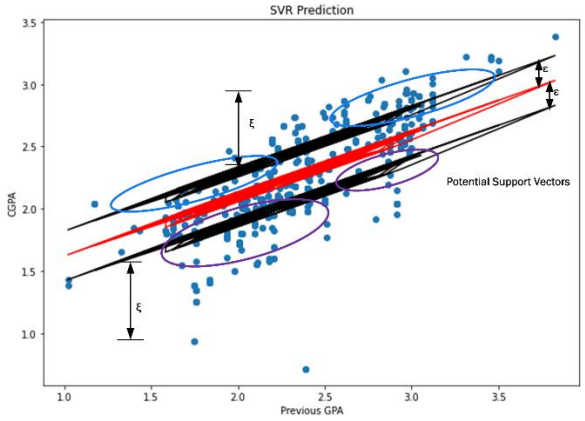

Figure 3: One-dimensional linear SVR (Previous GPA vs. CGPA)

The constraint equation is given as $|y_i - w_i x_i| \leq \epsilon$

The complete objective function with the regularization parameter is given as:

$$\underset{w}{\text{Min}} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} |\xi| \#(6)$$

Subject to constraints $|y_i - w_i x_i| \leq \epsilon + |\xi|$.

The regularization parameter determines the model's capacity to predict previously unobserved data, which manages the trade-off between attaining a low training error and testing error. The margin of error for values outside of ϵ increases as C rises, and vice versa. Equation 6 becomes equation 5 as C gets closer to zero.

3. RESULT AND DISCUSSION

The training and testing data were derived from the project dataset using an 80:20 split ratio. When modelling a regression problem, different metrics for evaluation and assessment of the performance of each of the models are deployed compared to evaluation metrics used in classification problems. Four regression evaluation metrics were used, namely, mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and R squared ($\boldsymbol{R^2}$).

The result of the evaluation of the linear regression and support vector regression models is presented in Table 2 below.

Table 2: Models' performance on the test dataset

Model Type	MAE	MAP	RMSE	$\boldsymbol{R^2}$
Linear Regression	0.195	0.0711	0.266	0.607
Support Vector Regression	0.212	0.074	0.272	0.601

The linear regression and support vector models performed similarly, predicting student CGPA based on feature variables outlined in Table 2. Similar R^2 scores (0.6) were obtained for both models, indicating that 60% of the variation of the CGPA score received by a student is explained by the input variables used in developing the models.

3.1 Explanatory Model of Linear Regression

The link between the result and the predictor variables can also be explained by the linear regression model (CGPA). Here, understanding the general connection between the variables is more important than the model's predictive power. The following conclusions may be drawn from Figure 4 below:

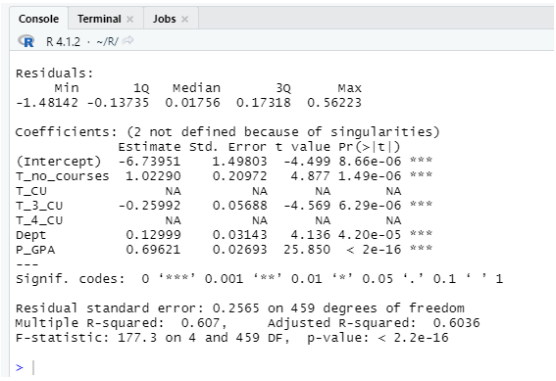

Figure 4: Summary of Linear Regression Model

i. **Intercept:** The intercept b_0 was obtained to be -6.739 and has a significance code of 0 (***), implying that the intercept has an inverse linear relationship with the predicted outcome (CGPA).

ii. **Total Number of Courses:** The total number of courses variable has a coefficient of 1.022. This variable has a significant 0 (***) code, signifying a direct relationship between the total number of courses and CGPA. Also, the value of 1.022 implies that for every unit increase in the total number of courses, the CGPA will be increased by 1.022.

iii. **Total Credit Unit:** There was no statistical significance between the entire credit unit and CGPA. This aligns with the fact that there was a slight disparity between the total credit units in all the four departments considered.

iv. **Total Number of Courses with more than three credit units:** There was a statistical significance between the total number of courses with three credit units. The significance with code 0 (***) indicates a robust inverse relationship between the variable and the response outcome. Any increase in the total number of courses with three credit units will decrease the CGPA by 0.259.

v. **Previous GPA:** The prior GPA has a strong correlation with the CGPA. This is indicated by a significant code of 0 (***). An increase in the unit of the previous GPA will increase the CGPA by 0.696.

The overall model had a p-value of $2.2e-16$, signifying a robust linear relationship between the predictor and response variables. Also, the R-squared value in Figure 1.6 showed a value of 0.607. This value indicated that the input predictor variables explained a 60% variation of the output variable (CGPA).

4. CONCLUSION

The fundamentals of accessing student academic performance have been examined in this study. Similar studies in this field have mostly excused the peculiarity of Nigeria's educational system. Additionally, it has been determined that although other studies have been conducted to evaluate student performance, the nature of their metrics,

the types of their parameters, and their variables are simply inapplicable to Nigeria's current academic system. Therefore, the requirement for a comprehensive and holistic approach that was simple to create without adding complexity and entirely valid for the educational system became evident very quickly. This model was designed, developed, and used based on this concept. This study made significant progress in predicting the cumulative grade point average.

The choice of implementing two models allowed for the robustness and validity of the research work. By using two models, it is possible to conclude that the work is valid and transcends beyond the limitation of a particular machine learning algorithm. Linear regression and support vector regression techniques were used. These two models performed well in their ability to predict the Student's CGPA. However, the Linear regression-based model performed slightly better than the support vector regression-based model. This further underscores that linear regression-based models were ideally suitable for regression problems. Also, using a linear regression-based model offers some advantages over support vector regression in areas such as its simplicity and ability to be used for model explanation and prediction purposes. Using the linear regression model, it was observed that four out of six variables had some form of linear relationship with response output. On the explanatory part of the linear regression-based model, it was concluded that the total number of courses offered in the semester, the type of department and previous grade point average had a positive linear relationship with the CGPA, implying a higher CGPA score. On the other hand, having a higher number of courses with three credit units affected the CGPA score negatively.

5. REFERENCES

[1] K. K. Obasi. **Teacher performance appraisal and quality secondary education delivery: The planning challenges**, *African Journal of Educational Research and Development*, vol. 4, no. 2, pp. 217–223, 2011.

[2] K. Ofori-Bruku. **The role of technical vocational education in Africa's economic development: Are the polytechnics still relevant?**, in *Proceedings of CAPA seminar*, 2005, pp. 32–42.

[3] S. F. Hamilton and M. A. Hamilton. **School, work, and emerging adulthood**, in *Emerging adults in America: Coming of age in the 21st century*, Washington: American Psychological Association, 2006, pp. 257–277.

[4] M. Kamal and A. Bener. **Factors contributing to school failure among school children in very fast developing Arabian Society**, *Oman Med. J.*, vol. 24, no. 3, pp. 212–217, Jul. 2009.

[5] C. Eleby. ***The Impact of a Student's Lack of Social Skills on their Academic Skills in High School. Online Submission.*** Online Submission, 2009.

[6] S. Hassel and N. Ridout. **An investigation of first-year students' and lecturers' expectations of university education**, *Front. Psychol.*, vol. 8, Jan. 2018.

[7] O. Tsolou and T. Babalis. **The contribution of family factors to dropping out of school in Greece**, *Creat. Educ.*, vol. 11, no. 08, pp. 1375–1401, 2020.

[8] E. A. Ekubo. **Data collection experience on educational data mining in Nigeria**, *Am J Comp Sci Inform Technol*, vol. 7, no. 2, 2019.

[9] G. E. Okereke. **A machine learning-based framework for predicting student's academic performance**, *Phys Sci Biophys J*, vol. 4, no. 2, 2020.

[10] H. Almarabeh and King Saud Bin Abdulaziz University for Health Sciences College of Science and Health Professions Riyadh, Kingdom of Saudi Arabia. **Analysis of students' performance by using different data mining classifiers**, *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 8, pp. 9–15, Aug. 2017.

[11] A. Zohair and L. Mahmoud. **Prediction of student's performance by modelling small dataset size**, *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1–18, 2019.

[12] E. T. Lau, L. Sun, and Q. Yang. **Modelling, prediction and classification of student academic performance using artificial neural networks**, *SN Appl. Sci.*, vol. 1, no. 9, Sep. 2019.

[13] N. Padhy. **The survey of data mining applications and feature scope**, *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 2, no. 3, pp. 43–58, Jun. 2012.

[14] C. Romero and S. Ventura. **Data mining in education**, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, Jan. 2013.

[15] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman. **Understanding the role of digital technologies in education: A review**, Sustainable Operations and Computers, vol. 3, pp. 275–285, 2022.