

PREDICTING STUDENTS' ACADEMIC PERFORMANCE USING ARTIFICIAL NEURAL NETWORK

Otse Moses Okpekwu¹, Rapheal Umoru Sunday PhD² & Peter Anyanwu C. PhD³

¹Federal College of Agriculture, Ibadan, otseokpekumoses@gmail.com/+2348064462253

²Wesley University Ondo, raphaelumoru@gmail.com/+2348052421067

³Wesley University Ondo, peter.anyanwu@wesleyuni.edu.ng/+2349150640833

ARTICLE INFO

Article No.: 069

Accepted Date: 25/08/2025

Published Date: 10/10/2025

Type: Research

ABSTRACT

University electronic learning (e-learning) has witnessed phenomenal growth, especially in 2020, due to the COVID-19 pandemic. This type of education is significant because it ensures that all students receive the required learning. The statistical evaluations are limited in providing good predictions of the university's e-learning quality. That is forcing many universities to go to online and blended learning environments. This paper presents an approach of statistical analysis to identify the most common factors that affect the students' performance and then use artificial neural networks (ANNs) to predict students' performance within the blended learning environment of Saudi Electronic University (SEU). Accordingly, this study generated a dataset from SEU's Blackboard learning management system. The proposed model's performance was evaluated through different statistical tests, such as error functions, statistical hypothesis tests, and ANOVA tests. The student's performance can be tested using a set of factors: the studying (face-to-face or virtual), percentage of attending live lectures, midterm exam scores, and percentage of solved assessments. The results showed that the four factors are responsible for academic performance. After that, we proposed a new ANN model to predict the students' performance depending on the four factors. Firefly Algorithm (FFA) was used for training the ANNs.

Keywords: students' performance; e-learning; artificial neural network; optimization algorithms; firefly algorithm

Introduction

E-Learning is an educational strategy that uses telecommunication technology to deliver information for education and training. It is defined as the ability of a system to transfer, manage, support, and supervise learning materials electronically. E-learning encompasses more than just computer-based teaching through the Internet or online information transmission. The two key elements of e-learning are electronic technology and the learning experience, respectively. The term “e-Learning” refers to both the technology and the method of instruction. In the last few years, we have seen an upward growth in the use of e-learning methods. This has been possible because of the availability of digital channels and the cost-effectiveness of e-learning. The main advantages of e-learning, such as the absence of constraints (in terms of date, room, duration of training), teacher availability, and cost-effectiveness in course delivery and management, prompted educational and training institutions to adopt e-learning by implementing an expanding list of technology-enabled platforms. According to a systematic study by the World Health Organization, 29% of studies evaluating knowledge improvement and 40% of studies evaluating skills improvement showed the benefit of e-learning compared to traditional learning. Because of the sudden outbreak of the COVID-19 pandemic worldwide, the education system across the globe is forced to switch from traditional learning systems to e-learning systems. UNESCO recommended that educational institutes gear up themselves with online learning tools. Due to COVID-19, e-learning strategies have become more popular and are now the most accessible means for education. However, with the successful usage of e-learning systems, there are some challenges facing the e-learning process, such as content transmission and delivery as well as enabling technologies. From the point view of Almaiah et al., student’s acceptance of e-learning is considered as one of the main challenges for the success of e-learning systems, while a study conducted by Al-Arabi et al. considers the technical issues as the main criteria for the success of e-learning system.

The blended learning system is an educational approach that combines e-learning and teaching face to face. Currently, blended learning is the most popular teaching method adopted by educational institutions due to its provision of flexible, timely, and continuous learning. In addition, blended learning increases the interaction between teachers and their students. The importance and benefits of a blended learning approach to improving teaching and learning are demonstrated in various studies, and many researchers consider it the “new normal”. The increasing popularity of blended learning has created a demand for improving the efficiency of this type of learning. Current research focuses on improving the quality of blended learning by using accurate prediction mechanisms to predict various aspects of blended learning efficiency, such as quality, probability of successful completion, the satisfaction of the participants in the training, detection of the learning styles, etc. Both management and educators benefit from the universities’ or schools’ effective development and evolution of the intervention plans, which are made possible by student performance prediction at the entry-level and subsequent periods. Several studies in the literature have addressed issues related to e-learning prediction and forecasting. Ioanna et al., proposed a student achievement prediction method to dynamically predict students’ final achievement and cluster them in two virtual groups using multiple feed-forward ANNs according to their performance. Using Fuzzy C-Means as a clustering algorithm, El Aissaoui et al. proposed a generic approach for detecting learning styles automatically according to a given learning styles model. Another study by Vilorio et al. models student engagement with the study material using data mining approaches to enhance e-learning processes. They used prediction rules whose interpretation will detect the educational process’s weaknesses

and evaluate the study material's quality. For improving learning systems in an e-learning environment, Novais et al. Proposed an intelligent learning system to monitor the patterns of students' behavior during e-assessments, to support the teaching procedure within school environments. There are opportunities to investigate the effect of technology in raising student grades in a mixed learning environment where ANNs and statistical analysis are merged.

The parallel and iterative ANN, which uses small processing units called neurons, is a technique. While the perceptron, a web of primary neurons, is the basic unit of a multilayer neural network. This paper describes a new ANN model using many factors to predict early the final grades of the students. Multilayer perceptron neural networks (MLPNNs) were chosen for their efficiency performance in predicting, and for their generalizability.

In addition, we used the nature-inspired metaheuristic algorithms to train ANN models. In parallel, the effectiveness of new factors for predicting student scores that made up the data set of this study was tested. In current years, metaheuristic optimization algorithms have become very popular in artificial intelligence (AI), optimization, engineering applications, data mining, and machine learning. These metaheuristic algorithms are now one of the most widely used techniques for optimization. They have several advantages over conventional techniques. The two most advantages are flexibility and simplicity. Metaheuristic algorithms are highly flexible and can deal with diverse objective functions, either discrete, continuous, or mixed. Moreover, these algorithms are simple to apply because of their ability to solve complex problems. They solve several real-world optimization problems in different domains, such as engineering, AI, and operation research. Various metaheuristic optimization algorithms such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and FFA have been used to find the optimal parameters of the ANNs models.

In this study, the FFA is applied as an optimization method to enhance the performance of the ANNs models. Note that, ANNs have effective, robust, and prominent features in capturing the nonlinear relationships between parameters and response in a complex system. FFA has proven that this algorithm has a high ability to search for the global optimum solution. Employing FFA and MLPNNs as tools to predict and improve the Students' Academic Performance will be very interesting and effectiveness. As a result, an integration of MLPNN with FFA has been undertaken in order to construct a hybrid prediction model for students' academic performance in blended learning modeling.

Research Objectives

1. To develop an Artificial Neural Network (ANN) model capable of predicting students' academic performance using relevant academic and personal data.
2. To identify and evaluate the key factors (e.g., previous academic records, attendance, socioeconomic status) that influence students' academic outcomes.
3. To assess the accuracy and performance of the ANN model in comparison with traditional statistical prediction methods.
4. To examine how the predictive model can be use as an early intervention tool to improve educational planning and student support.

Research Questions

1. What is the accuracy of Artificial Neural Networks in predicting students' academic performance?
2. What are the most significant input variables influencing students' academic performance?

3. How does the performance of ANN compare with traditional prediction models like linear regression?
4. To what extent can the ANN model be used to support early interventions in education?

Hypotheses

The following hypotheses were tested at 0.05 level of significance:

- H1.** There is no effect of attendance on final exams at the significance level ($\alpha = <_0.05$).
- H2.** There is no effect of assignments on final exams at the significance level ($\alpha = <_0.05$).
- H3.** There is no effect of mid-exams on final exams at the significance level ($\alpha = <_0.05$).
- H4.** There is no effect of virtual/face on final exams at the level of significance ($\alpha = <_0.05$).

Literature Review

Sayed and Baker introduced an ANN model as a type of supervised learning. They explained how to employ ANNs to create a convergent mathematical model using e-learning interactions and social analytics. Over eight years, Paul and Jefferson found that a more successful teaching strategy had been proven. They looked at score variations between genders and classifications to ascertain whether a certain teaching method had a larger influence on particular groups. There was no discernible difference in student performance between face-to-face and online students. Their results show that, regardless of gender or class position, environmental science concepts may be similarly translated for non-STEM majors in both conventional and online platforms. In their approach, Lau et al. combined traditional statistical analysis with ANN modeling and prediction of student performance. As the backpropagation training rule, the Levenberg–Marquardt algorithm is used. Through the use of statistical techniques, the performance of the neural network model is assessed. Despite its drawbacks, the neural network model has a decent overall prediction accuracy of 84.8%. In order to choose high-influence attributes with student performance, Khasanah and Harwati used feature selection.

They used Bayesian networks and decision trees, comparing the results to determine which provided the best predictions. The results demonstrated that student attendance and grade point average in the first semester was in the top rank of all feature selection methods, and Bayesian Network outperformed Decision Tree since it has a greater accuracy rate. Maheswari and Preethi employed deep learning to improve the placement performance of the student. The deep learning approach might influence and lead to the welfare of teachers, students, and educational institutions. To forecast the student's performance in placement, they used the Logistic Regression algorithm, and they looked at which algorithm provided the best accuracy for the dataset. Sultana et al. used the educational data mining technique to assess student performance. It predicts student performance and learning behavior by uncovering hidden knowledge from learning records or educational data. Using various classification systems, they investigated student performance and found the best one that produced the best results. Oyediji et al. utilized machine learning algorithms to examine the data collected from past student tests, including data on each student's age, family history, and attitude toward learning. They looked at linear regression for supervised learning, linear regression with deep learning and ANN, and training data with linear regression for supervised learning having the best mean average error. The cumulative grade point average (CGPA) upon graduation was used by Ibrahim and Rusli to gauge academic success. ANNs, decision trees, and linear regression are the three prediction models they created. They demonstrated that all three models produced more than 80% accuracy and that ANNs perform better than the other two models. Mandal et al. determined the ANNs weights by minimizing function cost or error. To train an ANN, the FFA—a powerful metaheuristic optimization

technique—is used, inspired by the firefly’s natural movement toward greater light. The simulation results demonstrate the Firefly Optimization training process’s high computational efficiency. Based on literature review, this study mainly used a new ANN model (MLPNN with FFA) to analyze the predictability of student performance and verified the generality of the prediction through new factors, which are mid-exams, assignments, attendances, and virtual/face.

Mathematical Modeling

Firefly Algorithm

Swarm intelligence algorithms are kinds of heuristic algorithms that are inspired by nature-observed phenomena. Such algorithms are looking for optimal solutions in the research space by conducting a cooperative population search. The FFA is one of the practical swarm intelligence algorithms proposed by Yang et al. in 2022. The FFA and its variants have recently emerged as powerful instruments for solving several optimization problems in different disciplines, such as engineering optimization, machine learning, path planning, and production scheduling. The FFA simulates the fireflies’ flashing behavior and effectively determines the optimal solutions (both global and local solutions). Every natural firefly exhibits luminous behavior to attract other mates. The attractiveness of the fireflies depends on factors, including the light intensity that the fireflies emit, their size, and location. The idealized fireflies’ flashing behavior occurs through the following scenario. First, each of the fireflies is attracted to other bright fireflies. Second, the fireflies’ attraction is decided by the firefly’s brightness, and it depends on the value of an objective function. This means that any less bright firefly moves towards the brighter one, and if there is no brighter firefly, the firefly will move randomly. The updates of the fireflies will be iterated until reaching the tolerance from the optimum value, or the maximum number of iterations is conducted.

Research Methodology

The major goal of the current study is to predict the impact of the student’s activity on his final grade during the semester. Other goals of the study include exploring the effects of teaching in virtual or face-to-face teaching and factors affecting student achievement. The data used were collected from the results of students from SEU for several different years and for other subjects.

Table 1. Descriptive statistics.

Variables	Mean	Std. Dev.	Min.	Max.	Skewness		Kurtosis	
					Statistic	Std. Error	Statistic	Std. Error
Final exams	35	15.111	0	50	-1.051	0.159	0.141	0.317
Mid-exams	16.06	6.227	0	25	-0.674	0.159	1.389	0.317
Assignments	21.2	6.253	0	25	-2.349	0.159	4.729	0.317
Attendance	2.71	2.942	0	8	1.867	0.159	4.557	0.317
Virtual/Face	0.49	0.501	0	1	0.034	0.159	-2.016	0.317

Table 1 shows the descriptive statistics of the input variables (mid-exams, assignments, attendance, and virtual/face) and output variable (final exams). The number of observations is 234 students. In the output variable, the final exams have (Mean = 35, standard deviation = 15.11, the maximum number = 50, skewness = 1.051, and kurtosis = 0.141). In the input variables, the mid-exams have (Mean = 16.06, standard deviation = 6.22, the maximum number = 40, skewness = 0.674, and kurtosis = 0.159). Furthermore, the assignments have (Mean = 21.2, standard deviation = 6.25, the maximum number = 25, skewness = 2.349, and kurtosis = 4.729). Moreover, the

attendances have (Mean = 2.71, standard deviation = 2.942, the maximum number = 18, skewness = 1.867, and kurtosis = 4.557). Finally, the virtual/faces have (Mean= 0.49, standard deviation = 0.501, the maximum number = 1, skewness = 0.034, and kurtosis = 2.016).

Results and Discussions

Studying the dataset using various statistical analysis techniques is crucial to assess the student's performance. This step allows researchers to understand the dataset before applying the ANNs. Multicollinearity tests, OLS, Fixed effect, and Random effect regression are used to test the effect of the significant input variables on the students' performance. The input variables are (mid-exams, assignments, attendances, and virtual/face). To achieve our goal, we conducted a statistical analysis using R software given in the following:

Multicollinearity tests: Table 2 shows the correlation between input and output variables. The variables are removed because of multicollinearity between input variables. The "no multicollinearity" refers to the absence of perfect multicollinearity, which is an exact (non-stochastic) linear relation among the input values. According to the conducted tests, we removed some variables from input variables that are strongly related to other input variables. The result shows weak correlations between input variables because it is less than 50%.

Furthermore, there is a strong correlation between assignments and mid-exam. There are strong negative correlations between attendance and final exams.

Table 2. Correlation between inputs and outputs variables.

	Final Exam	Virtual/Face	Mid-Exam	Assignment	Attendance
Final exam	1	0.474	0.573	0.633	- 0.646
Virtual/Face		1	0.132	0.059	- 0.06
Mid-exam			1	0.494	- 0.476
Assignment				1	- 0.484
Attendance					1

Note: Signif. codes: '***' 0.01, '*' 0.05 (2-tailed).

We investigate our proposed model using three approaches: the OLS, the fixed effects, and the random effects. Table 3 shows how to input variable bias may result in inaccurate results.

To investigate the direct impact of the input factors on the output variable, the OLS analysis is conducted. In the OLS, the input variables (virtual/face, midterm exams, assignments) positively affect final exams with a significant level < 1%.

Therefore, we reject H2, H3, and H4. However, there are adverse effects between attendance on final exams, with a significant level < 5%. Therefore, we reject H1. The R-square is 70%, the adjusted R-square is 69.63%, and F-statistic is 134.6 with a significant level < 1%. The fixed effects is an estimator model that can be used to analyze panel data. It does not allow the lag of the output variables to be included as an input variable in the model. In this study, we divided our cross-sectional data into two groups based on virtual/face, and then the fixed and random effects were applied. In fixed effect, the input variables (virtual/face, midterm exams, assignments) positively affect final exams with a significant level < 1%. Therefore, we reject H2, H3, and H4. However, there is no effect between attendance on final exams, with a significant level < 5%. Therefore, we accept H1. The R-square is 64.2%, the adjusted R-square is 24.85%, and F-statistic is 49.76, with a significant level of less than 1%.

Table 3. OLS, fixed effect, and random effect

Variables	OLS		Fixed Effect		Random Effect	
	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Error
(Intercept)	10.1592	3.3144			9.7678	3.3421
Virtual/Face	12.4691	1.0986	9.3633	11.0740	12.4444	1.1963
Mid exam	0.4303	0.1104	0.4163	0.1518	0.4298	0.1096
Assignment	0.7639	0.1167	0.898	0.1700	0.7847	0.1169
Attendance	-1.6207	0.2622	-1.6325	0.4080	-1.6334	0.2643
Observations	234		234		234	
R-square	0.7015		0.6420		0.6923	
Adjusted R-square	0.6963		0.2485		0.6869	
F-statistic/Chisq	134.6000		49.7612		518.514	

Signif. codes: ‘*’ 0.01

The random-effects model, a variance components model, is used to analyze panel data. In contrast to the fixed effects model, the individual-specific impact is a random variable that is unrelated to the input variables in the random-effects model. The input variables (virtual/face, midterm exams, assignments) have a beneficial effect on final exams with a significant level $< 1\%$. Therefore, we reject H2, H3, and H4. However, there are adverse effects between attendance on final exams with a significant level $< 5\%$. Therefore, we reject H1. The R-square is 69.23%, the adjusted R-square is 68.69%, and F-statistic is 518.514 with a significant level $< 1\%$.

We conduct a Hausman test to compare the two fixed effects and random effects models, with the null hypothesis being that random effects are favored over fixed effects. The null hypothesis is not true; hence, this test checks if the distinct mistakes correspond with the regressors in the models. We utilize the fixed effects model if the p-value is significant (for instance, p-value 0.05). Otherwise, the random-effects model is applied.

Conclusions

This study aimed to improve the students’ academic performance in Blended Learning. The data is collected from student activities at SEU. The statistical analysis successfully proved that there are sufficient effects of selected factors on the students’ performance. Multicollinearity tests, OLS, Fixed effect, and Random effect regression are used to select sufficient factors. The four effective factors are mid-exams, assignments, attendances, and virtual/face. After that, we proposed successfully an MLPNN model that effectively predicts student achievement. The structure of our model has four input neurons, ten hidden neurons, and one output neuron. The input neurons represent the four factors, and the output neuron represents final exams. In parallel, the optimal values of parameters (the input and output weights) for the MLPNN model were calculated successfully using FFA. Note that FFA has an excellent convergence rate and an intense exploration ability. Therefore, the hybrid MLPNN-FFA models are highly robust in prediction accuracy. Therefore, we recommend researchers to use the hybrid MLPNN-FFA models in their future studies. The proposed MLPNN model predicted the student’s performance in blended learning, to improve the learning process, and to reduce academic failure rates. It can also help administrators to enhance the blended learning system results.

References

- Al-araibi, A. A. M., Mahrin, M. N. R. B., & Yusoff, R. C. M. (2019). Technological aspect factors of e-learning readiness in higher education institutions: Delphi technique. *Education and Information Technologies*, 24(1), 567–590. <https://doi.org/10.1007/s10639-018-9780-9>
- Almaiah, M. A., Al-Khasawneh, A., & Althunibat, A. (2020). Exploring the critical challenges and factors influencing the e-learning system usage during COVID-19 pandemic. *Education and Information Technologies*, 25(6), 5261–5280. <https://doi.org/10.1007/s10639-020-10219-y>
- Crawford, J., Butler-Henderson, K., Rudolph, J., Malkawi, B., Glowatz, M., Burton, R., Magni, P., & Lam, S. (2020). COVID-19: 20 countries' higher education intra-period digital pedagogy responses. *Journal of Applied Learning & Teaching*, 3(1), 1–20.*
- Gabai, S., & Ornager, S. (2012, June 14). *Smart e-learning through media and information literacy*. In *Proceedings of the Fourth TCU International e-Learning Conference: Smart Innovations in Education & Lifelong Learning* (Bangkok, Thailand).
- Hu, P. J.-H., & Hui, W. (2012). Examining the role of learning engagement in technology-mediated learning and its effects on learning effectiveness and satisfaction. *Decision Support Systems*, 53(4), 782–792. <https://doi.org/10.1016/j.dss.2012.05.014>
- Kaup, S., Jain, R., Shivalli, S., Pandey, S., & Kaup, S. (2020). Sustaining academics during COVID-19 pandemic: The role of online teaching-learning. *Indian Journal of Ophthalmology*, 68(7), 1220–1221. https://doi.org/10.4103/ijo.IJO_1241_20
- Masie, E. (2008). What is the meaning of the E in e-learning? In E. Biech (Ed.), *ASTD handbook for workplace learning professionals* (p. 227). Association for Talent Development.
- Olafsen, R. N., & Cetindamar, D. (2005). E-learning in a competitive firm setting. In *Innovations in Education and Teaching International* (Vol. 42, pp. 325–335). Taylor & Francis.
- Shailaja, J., & Sridaran, R. (2014, March 6–7). *Taxonomy of e-learning challenges and an insight to blended learning*. In *Proceedings of the 2014 International Conference on Intelligent Computing Applications* (pp. 310–314). Coimbatore, India.
- Sun, P.-C., Tsai, R. J., Finger, G., Chen, Y.-Y., & Yeh, D. (2008). What drives a successful e-learning? An empirical investigation of the critical factors influencing learner satisfaction. *Computers & Education*, 50(4), 1183–1202. <https://doi.org/10.1016/j.compedu.2006.11.007>
- Wolbrink, T. A., & Burns, J. P. (2012). Internet-based learning and applications for critical care medicine. *Journal of Intensive Care Medicine*, 27(5), 322–332. <https://doi.org/10.1177/0885066611406857>