Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1561-1567
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Predicting Student Performance Using Deep Learning Models: A Comparative Study of MLP, CNN, BiLSTM, and LSTM with Attention

Gregorius Airlangga

Information System Study Program, Atma Jaya Catholic University of Indonesia, Indonesia

E-Mail: gregorius.airlangga@atmajaya.ac.id

Received Aug 18th 2024; Revised Sept 30th 2024; Accepted Oct 9th 2024
Corresponding Author: Gregorius Airlangga

Abstract

This study aims to predict student performance using deep learning models, including Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Long Short-Term Memory with Attention (LSTM with Attention). The dataset comprises student demographic and educational factors, and the models are evaluated using metrics such as MAE, RMSE, R^2, MSLE, and MAPE. The results show that the CNN model outperforms other models, achieving the highest accuracy in predicting student test scores. The MLP model also performs well, while the BiLSTM and LSTM with Attention models exhibit lower predictive performance. High MAPE values across models suggest a need for alternative metrics in future research. This study highlights the importance of selecting suitable model architectures for predictive tasks in education, emphasizing the effectiveness of convolutional layers in capturing complex patterns.

Keyword: Bidirectional LSTM, Convolutional Neural Networks (CNN), Deep Learning Models, Educational Data Analysis, Student Performance Prediction,

1. INTRODUCTION

The academic performance of students is a complex phenomenon influenced by multiple factors, including demographic characteristics, family background, and school-related factors [1]–[3]. Understanding these influences is crucial for developing effective educational interventions [4]. Despite considerable research efforts, the relationship between student performance and these multifaceted factors remains a topic of debate[5]. Traditional statistical methods have provided valuable insights but often fall short in capturing the nonlinear and intricate relationships within educational data [6]. With advancements in machine learning and deep learning, new opportunities arise to model these complexities more effectively, enabling a deeper exploration of how variables such as gender, ethnicity, parental education, lunch status, and test preparation impact student test scores [7]. Recent developments in artificial intelligence (AI) and deep learning have shown promise in various fields, including educational data mining [8]. Advanced models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been successfully employed to uncover patterns within large datasets, demonstrating superior predictive capabilities [9]. However, there remains a gap in the literature regarding the application of these advanced models to the educational domain, specifically in predicting student performance using diverse predictors [10].

Educational research has traditionally focused on socioeconomic status and family background as key determinants of student performance. Parental education, a significant component of this background, has been repeatedly shown to correlate with higher academic achievement [11]. Moreover, gender differences have been extensively examined, with findings suggesting that boys and girls exhibit varying strengths in subjects like mathematics, reading, and writing [12]. School-related factors, including lunch programs and test preparation courses, also play a role. Students with access to adequate nutrition and structured test preparation often exhibit improved academic performance [13]. While these studies provide a foundational understanding, they often employ traditional statistical techniques, limiting their ability to uncover complex interactions among predictors. Machine learning approaches, such as Decision Trees and Random Forest, have addressed some limitations by improving predictive accuracy and handling nonlinear relationships [14]. However, the literature lacks comprehensive analyses using deep learning models, particularly those equipped to handle sequential and high-dimensional educational data.

The increasing reliance on data-driven strategies in education underlines the urgent need for advanced analytical tools to interpret complex educational datasets [15]. Traditional methods provide limited insight into

the nuanced interplay of diverse factors affecting student performance [16]. With educational systems facing growing demands for personalized learning and targeted interventions, it is imperative to explore how deep learning models can enhance our understanding and prediction of student outcomes [17]. This urgency is driven by the potential of these models to inform policies and practices that can effectively support diverse student populations. Machine learning and deep learning have become central to educational data analysis [18]. Ensemble methods like Random Forest and Gradient Boosting have gained popularity for their robust performance in predictive tasks [19]. Deep learning models, including CNNs and RNNs, offer further advancements, capturing complex data patterns through hierarchical and sequential processing [20]. More recently, attention mechanisms have been introduced to enhance the interpretability of deep learning models by highlighting influential features in the decision-making process [21]. Despite their success in other domains, the application of these state-of-the-art models to the domain of student performance prediction remains limited.

Existing studies predominantly utilize traditional statistical and basic machine learning models to analyze student performance, often focusing on isolated predictors [22]. There is a lack of research that holistically applies deep learning models to account for complex interactions among a comprehensive set of predictors [23]–[25]. Additionally, the challenge of interpreting deep learning models within the educational context has not been sufficiently addressed. This research aims to fill these gaps by implementing and comparing various deep learning models, investigating their predictive capabilities, and examining how attention mechanisms can be employed to improve model interpretability. The goal of this research is to develop a comprehensive predictive framework for student performance using deep learning models. By integrating a diverse set of predictors, including demographic and educational variables, the study aims to evaluate the performance of advanced deep learning architectures Deep MLP, Deep CNN, Bidirectional LSTM, and LSTM with Attention in predicting student test scores. This framework not only seeks to improve predictive accuracy but also strives to enhance the interpretability of the models, providing actionable insights into the factors that most significantly impact student outcomes.

In addition, this study contributes to educational data mining by applying advanced deep learning models, including Deep CNN, Bidirectional LSTM, and LSTM with Attention, to the prediction of student performance. It offers a comparative evaluation of these models' effectiveness in capturing the complex relationships between student test scores and influencing factors. Additionally, this research explores the interpretability of these models, enhancing our understanding of how specific variables affect student performance. The article proceeds with a methodology section outlining the dataset, feature engineering, and model development, followed by an evaluation of the models' performance. The results section presents a detailed analysis of the predictive accuracy and interpretability of each model. In addition, we have add discussion section in order to explore the implications of the findings for educational practice and policy. Finally, the conclusion summarizes the contributions and suggests directions for future research.

2. MATERIALS AND METHOD

As presented in figure 1, this section provides an in-depth explanation of the dataset used, the preprocessing methods applied, the development of the deep learning models, and the evaluation techniques employed. The methodology is carefully structured to allow for reproducibility while ensuring a clear understanding of the mathematical and computational rigor involved.

2.1. Dataset Description

The dataset employed in this study consists of 1,000 records, each containing information about a student's demographics, parental background, and educational experiences. Each record includes features such as the student's gender, race/ethnicity, parental level of education, lunch status, and participation in a test preparation course. The dependent variables are the scores obtained by the students in three academic subjects: mathematics, reading, and writing. These scores are considered continuous variables and serve as the targets for the regression models. Formally, let (X) represent the matrix of input features with dimensions ($N \times d$), where ($N = 1000$) denotes the number of instances, and (d) is the number of features for each student. The target vector (Y) comprises the test scores for each student, structured as a matrix ($Y \in R^{N\times 3}$), where each column corresponds to a different subject score.

2.2. Data Preprocessing

To prepare the dataset for input into the deep learning models, several preprocessing steps were applied to ensure that the data is in a suitable format and scale. First, categorical variables including gender, race/ethnicity, parental education level, lunch status, and test preparation were transformed using one-hot encoding. One-hot encoding converts categorical values into a binary matrix representation, effectively allowing the model to process these non-numeric features. Mathematically, for a categorical feature (C) with (k) unique categories, one-hot encoding maps this feature to a binary vector space (R^k). Each category $(c_i) in (C)$ is transformed into a binary vector (e_i) where the (i)-th position is 1, and all other positions are 0.

This transformation ensures that categorical data is numerically represented in a way that does not imply any ordinal relationship between categories.

Next, continuous variables representing the student's scores in mathematics, reading, and writing were standardized. Standardization is critical when working with neural networks, as it ensures that each feature contributes equally to the model's learning process. The standardization process adjusts each feature to have a mean of zero and a standard deviation of one. Given a continuous variable (x), the standardized value (z) is computed as $z = \frac{x-\mu}{\sigma}$ where (μ) is the mean of (x) and (σ) is its standard deviation. This transformation results in a new dataset where each feature has a mean of zero and unit variance, preventing any single feature from dominating the learning process due to its scale.

Subsequently, the dataset was split into training and testing sets to facilitate model evaluation. The split ratio was set at 80:20, meaning 80% of the data was used for training the models, while the remaining 20% was reserved for testing their generalization capabilities. Let (X_{train}) and (X_{test}) denote the training and test sets, respectively. The splitting process is defined by a function (g), where $g(X) = (X_{\text{train}}, X_{\text{test}})$. This ensures that the models are trained on one subset of the data and evaluated on an unseen subset, providing a reliable estimate of performance.

For models such as CNNs and LSTMs, the input data needed to be reshaped to match the expected input dimensions. For CNNs, the input was reshaped into a three-dimensional tensor to exploit the convolutional operations effectively. Specifically, the reshaped input (X_{reshape}) is defined as $X_{\text{reshape}} \in R^{N\times t\times f}$ where (N) is the number of samples, (t) represents the sequence length, and (f) is the number of features. This restructuring is crucial for the CNN and LSTM models to capture spatial and sequential patterns in the data, allowing these architectures to learn more complex relationships within the dataset.

2.3. Model Development

The study involves developing and evaluating multiple deep learning models, each tailored to predict student performance based on the input features. These models include a Deep Multilayer Perceptron (MLP), a Convolutional Neural Network (CNN), a Bidirectional Long Short-Term Memory (BiLSTM) network, and an LSTM network with an Attention mechanism. The Deep MLP model is structured as a fully connected neural network with three hidden layers. Each hidden layer applies a transformation function to its input, followed by the ReLU activation function. Formally, the transformation in each hidden layer (l) is expressed as $h^{(l)} = \phi\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)$ where ($W^{(l)}$) and ($b^{(l)}$) are the weight matrix and bias vector for layer (l), ($h^{(l-1)}$) is the output from the previous layer, and \(\phi \) is the ReLU activation function defined as ($\phi(x) = \max(0, x)$). Dropout regularization is applied to each hidden layer, where a fraction of the neurons is randomly set to zero during training. This regularization technique is denoted as $h^{(l)}_{\text{drop}} = h^{(l)} \cdot \text{mask}$ where the mask is a binary vector with each element sampled from a Bernoulli distribution with a success probability of ($1 - p$), where (p) is the dropout rate. The output layer consists of a single neuron for regression, providing the predicted test score.

The Convolutional Neural Network (CNN) model consists of one-dimensional convolutional layers designed to detect local patterns within the input features. The convolutional operation applies a set of filters to the input, generating feature maps. Mathematically, this operation is defined as $(W * X)(t) = \sum_{\tau=1}^{k} W(\tau)X(t+\tau)$ where (W) is the filter (kernel) of size (k) and (X) is the input sequence. The output of the convolutional layers is passed through the ReLU activation function, followed by a flattening layer that converts the 3D output into a 1D vector. A fully connected layer with 128 neurons further processes this vector, and a dropout layer is applied to prevent overfitting. The final layer produces the predicted score for each student.

The Bidirectional LSTM model captures temporal dependencies in both directions within the input sequence. Standard LSTM networks process sequences in a single direction; however, Bidirectional LSTMs process the input sequence forward and backward, capturing relationships from both ends. The cell state (c_t) and hidden state (h_t) in an LSTM cell are computed using the input (x_t) at time (t) and the previous states $h_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1})$. The bidirectional wrapper aggregates these computations to form a richer representation of the sequence data, which is then passed to a dense layer for final prediction. The LSTM with Attention mechanism enhances the model's ability to focus on relevant parts of the input sequence. The attention mechanism assigns a weight (α_t) to each time step in the LSTM output, indicating its importance in the final decision-making process. The attention scores are calculated using a learned vector (u) and the hidden states (h_t) is $\alpha_t = \text{softmax}(u^T \cdot h_t)$. In addition, the context vector (c), representing the relevant information, is computed as a weighted sum of the hidden states as presented as $c = \sum_t \alpha_t h_t$. This context vector is then used as input to the final dense layer, allowing the model to make predictions based on the most influential parts of the sequence.

2.4. Cross-Validation

To ensure that the models are rigorously evaluated, 10-fold cross-validation was employed. This method divides the dataset into 10 equally sized subsets, or folds. In each iteration, one-fold is held out as the test set, and the remaining nine folds are used for training. This process is repeated 10 times, with each fold serving as the test set once. Let the dataset (X) be partitioned into (k) subsets, ($X = \{X_1, X_2, \dots, X_k\}$), where ($k = 10$). $For\ each\ iteration(i)$ {Train set $= \bigcup_{j \neq i} X_j$; Test set $= X_i$}. This approach provides a more reliable estimate of the model's performance by reducing variance caused by the specific train-test split and ensuring that every data point is used for both training and testing. To ensure optimal performance, each model underwent an extensive hyperparameter tuning process. For the CNN model, the kernel size, number of filters, and learning rate were tuned using a grid search approach. The LSTM model parameters, including the number of units in each layer and dropout rate, were fine-tuned to prevent overfitting. Specifically, a Bayesian optimization technique was employed to determine the optimal hyperparameter configurations. This approach provided a systematic way to balance model complexity with training efficiency, ensuring the reproducibility of results across different datasets and experimental settings

2.5. Evaluation Metrics

The performance of each model was evaluated using several metrics, capturing different aspects of the prediction quality. The Mean Absolute Error (MAE) measures the average magnitude of prediction errors and is defined as MAE $= \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|$ where (y_i) is the actual score, ($\widehat{y_i}$) is the predicted score, and \(N \) is the total number of instances. Furthermore, the Root Mean Squared Error (RMSE) penalizes larger errors more severely and is given by RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}$, This metric emphasizes larger deviations, making it sensitive to outliers in the data. The R-Squared (R^2) value, representing the proportion of variance in the dependent variable that is predictable from the independent variables, is calculated as $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$ where ($\bar{y}$) is the mean of the observed values. The Mean Squared Logarithmic Error (MSLE) assesses the ratio between predicted and true values, useful when predictions can vary by orders of magnitude as presented in MSLE $= \frac{1}{N}\sum_{i=1}^{N}(\log(1 + y_i) - \log(1 + \widehat{y_i}))^2$. Finally, the Mean Absolute Percentage Error (MAPE) provides a percentage-based measure of error as presented as MAPE $= \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$. These metrics provide a comprehensive evaluation of model performance, capturing both the accuracy and the nature of the errors. The combination of cross-validation and multiple evaluation metrics ensures a thorough assessment of the deep learning models in predicting student performance.

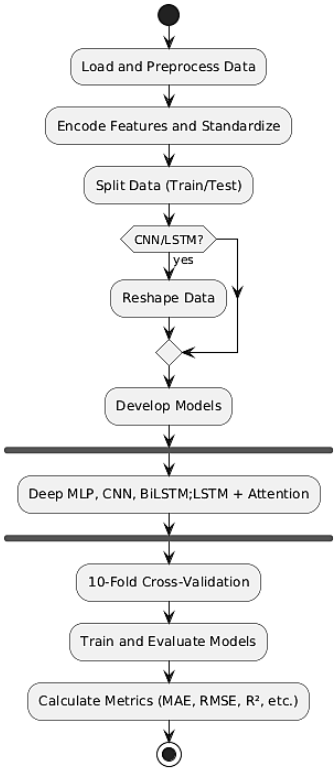

Figure 1. Research Methodology

3. RESULTS AND DISCUSSION

In this section, we present the performance results of the four deep learning models—Deep Multilayer Perceptron (MLP), Deep Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Long Short-Term Memory with Attention (LSTM with Attention). The models were evaluated using five metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), Mean Squared Logarithmic Error (MSLE), and Mean Absolute Percentage Error (MAPE). The performance across these metrics provides insight into the models' prediction accuracy and error distribution.

3.1. Result

As presented in the table 1, the Deep MLP model achieved an average MAE of 4.9704 with a standard deviation of 0.5546, indicating that the absolute difference between the predicted and actual scores is relatively low. The RMSE was 6.1817 ± 0.6641, slightly higher than the MAE due to the quadratic nature of RMSE, which penalizes larger errors more heavily. The R^2 value of 0.8286 ± 0.0365 shows that the Deep MLP model explains approximately 82.86% of the variance in the target variable, indicating a strong predictive ability. The MSLE value of 0.0213 ± 0.0324 shows that the logarithmic errors are minimal, which is expected when dealing with predictions that do not span several orders of magnitude. However, the MAPE value is unusually high, with an average of 1.11×10^{14}, which suggests that the percentage error is highly inflated. This anomaly is likely due to the presence of very small true values in the test set, leading to extremely large percentage differences when the predicted value differs slightly. This issue highlights the limitations of MAPE when dealing with very small target values.

The Deep CNN model outperformed the Deep MLP model in all metrics. The average MAE was 4.6900 ± 0.4084, and the RMSE was 5.8504 ± 0.4239, reflecting smaller errors overall. The R^2 value of 0.8472 ± 0.0242 indicates that the model explains a higher proportion of the variance (around 84.72%) compared to the MLP model. This suggests that the convolutional layers in the CNN architecture are effective at capturing important patterns in the data. The MSLE value of 0.0168 ± 0.0217 indicates that the CNN model makes fewer logarithmic errors compared to the MLP. Like the MLP, the CNN also suffers from high MAPE values (6.15×1013), which are likely caused by the same issue of small true values resulting in disproportionately high percentage errors. Nevertheless, the lower MAE, RMSE, and higher R^2 values indicate that the Deep CNN is better suited for this task.

The Bidirectional LSTM model's performance was less favorable compared to the MLP and CNN. The average MAE of 6.3668 ± 1.2688 indicates that the model made larger errors in its predictions. The RMSE of 7.8393 ± 1.3467 reinforces this finding, showing that the model penalizes larger errors more heavily, leading to a less accurate performance overall. The R^2 value of 0.7091 ± 0.1323 shows that the Bidirectional LSTM model explains approximately 70.91% of the variance, which is a significant drop compared to the MLP and CNN models. This suggests that while the BiLSTM model captures some temporal patterns, it does not generalize as well as the other models. Additionally, the MSLE value of 0.0240 ± 0.0136, although low, is slightly higher than that of the MLP and CNN models, further indicating that the BiLSTM model struggles with small prediction errors. Interestingly, the MAPE value of 3.85×1013, while still high, is lower than both the MLP and CNN models, suggesting that the BiLSTM model performs relatively better in minimizing percentage errors, even though its overall accuracy is lower.

The LSTM with Attention model performed the worst across all metrics. The MAE of 13.9078 ± 0.8064 indicates that this model made significantly larger errors in its predictions compared to the other models. Similarly, the RMSE of 17.2022 ± 0.9344 reflects that the model heavily penalizes large errors, resulting in poor overall performance. The R^2 value of -0.3097 ± 0.0972 shows that the model does not explain the variance in the data and may even predict worse than a simple mean prediction model. The MSLE of 0.0906 ± 0.0555, though relatively high, is still within an acceptable range given the large errors produced by the model. However, the MAPE value of 2.61×10^{14} further reinforces the model's inadequacy in making accurate predictions. This poor performance suggests that the addition of an attention mechanism may have introduced complexity that the model was unable to handle effectively, resulting in significant overfitting and poor generalization to unseen data.

Table 1. Comparison Results of Deep Learning Variants

Model	MAE	RMSE	R^2	MLSE	MAPE
Deep MLP	4.9704 ± 0.5546	6.1817 ± 0.6641	0.8286 ± 0.0365	0.0213 ± 0.0324	1.11×10^{14}
Deep CNN	4.6900 ± 0.4084	5.8504 ± 0.4239	0.8472 ± 0.0242	0.0168 ± 0.0217	6.15×10^{13}
Bidirectional LSTM	6.3668 ± 1.2688	7.8393 ± 1.3467	0.7091 ± 0.1323	0.0240 ± 0.0136	3.85×10^{13}
LSTM with Attention	13.9078 ± 0.8064	17.2022 ± 0.9344	-0.3097 ± 0.0972	0.0906 ± 0.0555	2.61×10^{14}

3.2. Discussion

The results indicate that the Deep CNN model outperformed the other models in predicting student performance, as evidenced by its superior MAE, RMSE, and R^2 scores. The convolutional layers likely allowed the model to effectively capture hierarchical patterns within the input features, leading to better generalization on unseen data. The Deep MLP model also performed well, albeit slightly worse than the CNN model, while the Bidirectional LSTM model demonstrated that temporal dependencies did not provide significant improvements for this task. The LSTM with Attention model's poor performance suggests that the added complexity of an attention mechanism may not always be beneficial, particularly in datasets where temporal or sequential relationships are not as strong. Instead of improving performance, the attention mechanism may have introduced unnecessary noise into the learning process, leading to overfitting. It is also worth noting the abnormally high MAPE values across all models, particularly in the MLP and CNN models. This suggests that MAPE may not be a suitable metric for this task, as it tends to overinflate errors when the true values are small. Future work could explore alternative error metrics that are more robust to small target values, such as symmetric mean absolute percentage error (sMAPE). Although the models performed well, a more in-depth analysis of each model's limitations is necessary. For instance, the CNN model may overfit the data due to its complex architecture, whereas the LSTM with Attention model exhibited poor performance likely due to the added complexity that was not beneficial for this particular dataset. Additionally, the BiLSTM struggled to generalize due to a lack of temporal dependencies in the data. Moreover, biases inherent in the dataset, such as imbalanced demographic representation, could have influenced the model outcomes, leading to skewed predictions.

4. CONCLUSION

This study investigated the performance of four deep learning models—Deep Multilayer Perceptron (MLP), Deep Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Long Short-Term Memory with Attention (LSTM with Attention)—in predicting student test scores based on demographic and educational factors. The models were evaluated using multiple metrics, including MAE, RMSE, R^2, MSLE, and MAPE, to provide a comprehensive view of their predictive capabilities. The results indicate that the Deep CNN model consistently outperformed the other models, achieving the lowest MAE and RMSE and the highest R^2, suggesting that its convolutional architecture was effective in capturing relationships within the dataset. The Deep MLP also performed well, though slightly less accurate than the CNN model. The Bidirectional LSTM model, which aimed to capture temporal dependencies, did not show significant improvements and lagged behind the MLP and CNN in accuracy. Finally, the LSTM with Attention model, which was expected to improve performance through its attention mechanism, performed poorly, indicating potential overfitting and an inability to handle the complexity of the task.

One notable observation across all models was the abnormally high MAPE values, which indicate that this metric may not be suitable for datasets where small true values exist. These high percentage errors suggest that alternative metrics such as sMAPE or RMSLE may be more appropriate for future studies dealing with similar datasets. In conclusion, this research underscores the importance of selecting the right model architecture for predictive tasks. Convolutional layers in deep learning models, as seen in the CNN, can capture complex patterns and yield better generalization. However, not all advanced models, such as LSTM with Attention, necessarily improve performance, especially when the task does not heavily depend on sequential or temporal relationships. Future research should consider expanding the dataset by including additional features, such as student behavior metrics or teacher evaluations, to provide a more comprehensive analysis. Exploring other advanced machine learning techniques, such as Transformer models or hybrid ensemble methods, could also improve the robustness of the findings. Furthermore, the integration of interpretability techniques, like SHAP values or LIME, would allow for a deeper understanding of model predictions, thus aiding in their practical application in personalized education.

REFERENCES

[1] A. Costa *et al.*, "Determinants of academic achievement from the middle to secondary school education: A systematic review," *Soc. Psychol. Educ.*, pp. 1–40, 2024.

[2] A. A. Adongo, J. M. Dapaah, and D. Wireko, "The influence of family size on academic performance of high school students in Ghana," *SN Soc. Sci.*, vol. 2, no. 9, p. 179, 2022.

[3] D. W. Ambaye, "Determinants of academic achievement among grade ten students at Menkorer secondary school, Ethiopia: the role of individual, familial and school characteristics," *Cogent Educ.*, vol. 11, no. 1, p. 2299523, 2024.

[4] D. H. Bailey, G. J. Duncan, F. Cunha, B. R. Foorman, and D. S. Yeager, "Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions," *Psychol. Sci. Public Interes.*, vol. 21, no. 2, pp. 55–97, 2020.

[5] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *Ieee Access*, vol. 8, pp. 203827–203844, 2020.

[6] F. Harrou, Y. Sun, A. S. Hering, M. Madakyaru, and others, *Statistical process monitoring using advanced data-driven and deep learning approaches: theory and practical applications*. Elsevier, 2020.
[7] A. E. Schwartz and M. W. Rothbart, "Let them eat lunch: The impact of universal free meals on student performance," *J. Policy Anal. Manag.*, vol. 39, no. 2, pp. 376–410, 2020.
[8] A. Bozkurt, A. Karadeniz, D. Baneres, A. E. Guerrero-Roldán, and M. E. Rodr\'\iguez, "Artificial intelligence and reflections from educational landscape: A review of AI Studies in half a century," *Sustainability*, vol. 13, no. 2, p. 800, 2021.
[9] M. Zulqarnain, R. Ghazali, M. G. Ghouse, Y. M. M. Hassim, and I. Javid, "Predicting financial prices of stock market using recurrent convolutional neural networks," *Int. J. Intell. Syst. Appl.*, vol. 13, no. 6, p. 21, 2020.
[10] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interact. Learn. Environ.*, vol. 31, no. 6, pp. 3360–3379, 2023.
[11] J. Ogg and C. J. Anthony, "Process and context: Longitudinal effects of the interactions between parental involvement, parental warmth, and SES on academic achievement," *J. Sch. Psychol.*, vol. 78, pp. 96–114, 2020.
[12] B. M. Casey and C. M. Ganley, "An examination of gender differences in spatial skills and math attitudes in relation to mathematics success: A bio-psycho-social model," *Dev. Rev.*, vol. 60, p. 100963, 2021.
[13] R. R. Weaver *et al.*, "University student food insecurity and academic performance," *J. Am. Coll. Heal.*, vol. 68, no. 7, pp. 727–733, 2020.
[14] L. Li, S. Rong, R. Wang, and S. Yu, "Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review," *Chem. Eng. J.*, vol. 405, p. 126673, 2021.
[15] D. Papadopoulos and M. M. Hossain, "Education in the age of analytics: maximizing student success through big data-driven personalized learning," *Emerg. Trends Mach. Intell. Big Data*, vol. 15, no. 9, pp. 20–36, 2023.
[16] R. Deng, P. Benckendorff, and D. Gannaway, "Linking learner factors, teaching context, and engagement patterns with MOOC learning outcomes," *J. Comput. Assist. Learn.*, vol. 36, no. 5, pp. 688–708, 2020.
[17] S. Maghsudi, A. Lan, J. Xu, and M. van Der Schaar, "Personalized education in the artificial intelligence era: what to expect next," *IEEE Signal Process. Mag.*, vol. 38, no. 3, pp. 37–50, 2021.
[18] M. Ya\ugc\i, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, p. 11, 2022.
[19] G.-W. Cha, H.-J. Moon, and Y.-C. Kim, "Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables," *Int. J. Environ. Res. Public Health*, vol. 18, no. 16, p. 8530, 2021.
[20] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Comput. Sci.*, vol. 2, no. 6, p. 420, 2021.
[21] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, 2022.
[22] Y. T. Badal and R. K. Sungkur, "Predictive modelling and analytics of students' grades using machine learning algorithms," *Educ. Inf. Technol.*, vol. 28, no. 3, pp. 3027–3057, 2023.
[23] M. Pichler and F. Hartig, "Machine learning and deep learning—A review for ecologists," *Methods Ecol. Evol.*, vol. 14, no. 4, pp. 994–1016, 2023.
[24] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," *Nat. Mach. Intell.*, vol. 2, no. 9, pp. 500–508, 2020.
[25] J. Devaraj, R. Madurai Elavarasan, G. M. Shafiullah, T. Jamal, and I. Khan, "A holistic review on energy forecasting using big data and deep learning models," *Int. J. energy Res.*, vol. 45, no. 9, pp. 13489–13530, 2021.