ISSN No:-2456-2165

Students Performance Prediction

Dona Boby[1], Megha Madhu[2], Ann Mary Danty[3]
[1,2,3]Department of Mathematics, Amrita Vishwa Vidhyapeetham, Kochi, India

Abstract:- Abstract Student performance prediction is an important aspect of education that has gained significance in recent years. Predicting the academic outcomes of students can help educators identify students who are at risk of falling behind and provide them with targeted interventions to improve academic performance. New technologies such as deep learning have revolutionized the way student performance prediction is done. Deep learning algorithms can analyze large amounts of data and identify patterns that would be difficult to detect using traditional statistical methods. In the proposed study, the dataset of students in Portuguese school contains various features such as age, gender, family background, study time, travel time, weekly study time, etc. The deep learning techniques employed in this study include Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), Convolutional Neural Network(CNN) and Bi-directional LSTM. The performance of these deep learning models was evaluated using metrics such as accuracy, mean squared error (MSE), and mean absolute error (MAE). This study demonstrates the effectiveness of deep learning techniques in predicting student performance and can be used as a basis for developing interventions to improve academic outcomes.

Keywords:- *Deep Learning;Academic Performance;Early Prediction.*

I. INTRODUCTION

Education plays a crucial role in shaping the future of children and society as a whole. It provides children with the knowledge and skills they need to succeed in life and make positive contributions to their communities. Education also helps children develop critical thinking, problemsolving, and decision-making abilities that will serve them well throughout their lives. Furthermore, education can promote social mobility, reduce poverty and inequality, and promote economic growth and development [1] . Overall, investing in education for children is essential for building a brighter, more prosperous, and equitable future for all.

Education is the process of acquiring knowledge, skills, values, and attitudes through various methods such as teaching, training, research, and practical experience. Education can take place in formal settings such as schools and universities, as well as informal settings such as workplaces and homes. It is not just about memorizing facts and figures, but also about developing social skills, creativity. Education helps individuals to broaden their perspectives and understand diverse cultures, which in turn promotes tolerance and mutual respect. Moreover, education is a fundamental human right that should be accessible to all, regardless of their background, gender, or socioeconomic status.

Education has undergone a significant transformation with the widespread application of technology in recent years. Technology has allowed for the development of new teaching methodologies and learning tools that cater to diverse learning styles and promote greater student engagement. From online courses to virtual classrooms, technology has made education more accessible and convenient than ever before, and it is expected to continue to play a crucial role in the future of education. However, it is important to ensure that the integration of technology in education is done thoughtfully and with a focus on enhancing learning outcomes, rather than simply replacing traditional teaching methods.

Educational data mining is a field that involves analyzing large sets of educational data to identify patterns, trends, and insights that can inform instructional decision-making [2]. This includes data from a range of sources, such as student assessments, attendance records, and demographic information. By analyzing this data, educators can gain a better understanding of student performance and behavior, and develop targeted interventions to improve outcomes. Our study demonstrates that the use of deep learning models can significantly improve the accuracy of predicting students' academic performance compared to traditional statistical models.

II. RELATED WORK

Over the past few years, there have been numerous studies aimed at discovering various patterns and strategies that can enhance students academic performance. Different methodologies and tools are used to visualize and analyze the data. Some of related works that have been done so far discussed on this section:

A. Nabil, M. Seyam and A. Abou-Elfetouh [3] presented a study to explore the efficiency of deep learning in the field of Educational Data Mining, for predicting students academic performance, inorder to identify the students who are at a risk of failure. The study used a public 4-year university dataset and developed predictive models to forecast students' performance in upcoming courses based on their grades in the previous courses of the first academic year. Various models, including a deep neural network (DNN), decision tree, random forest, gradient boosting, logistic regression, support vector classifier, and K-nearest neighbor, were employed for this purpose. The DNN model proposed in the study exhibited a high accuracy of 89 %, outperforming other models such as decision tree, logistic regression, support vector classifier, and K-nearest neighbor. The model was able to predict students' performance in the course data structure and thus helped to identify those students at the risk of failure in an early stage of a semester.A recent study focused on predicting student performance using the attention-based Bidirectional Long Short-Term Memory

(BiLSTM) network [4]. . They incorporated advanced feature classification and prediction techniques and found that the combination of BiLSTM and attention mechanism resulted in superior performance and achieved an accuracy of 90.16 %.

M. R. Islam Rifat, A. S. M. Badrudduza and A. Al Imran [5] have proposed a Deep Neural Network (DNN) based model to predict the final CGPA of undergraduate business students. They collected a real dataset by gathering transcript data from the marketing department of a reputable public university in Bangladesh. The dataset consisted of transcript data from students who graduated in 2013, 2014, 2015, and 2016. The researchers compared their proposed model's performance to a baseline decision tree model. The results showed that their DNN-based model significantly outperformed the baseline model. The proposed model reduced the mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) by 0.0146, 0.0431, and 6.043, respectively, compared to the baseline model.

Similarly, another study for finding Student Learning Outcomes in Learning Management Systems [6]. By combining two methods; namely, CNN to extract effective features from the data, and LSTM to identify the interdependence of data in time series data, the performance prediction accuracy was improved compared with state-of-the-art methods whereas testing data accuracy is 94.3 %.

A study in which an undergraduate database was used which is obtained from the Student Advisory and Support Center at Bina Nusantara University that comprises a total of 46,670 students enrolled from 2010 to 2017. The study introduced a dual-input deep learning model [7], capable of concurrently analyzing time-series and tabular data to forecast student GPA. The model proposed attained the most exceptional results among all assessed models, achieving a GPA prediction with a 4.0 scale, featuring a 0.4142 Mean Squared Error (MSE) and a 0.418 Mean Absolute Error (MAE).

A two-year long analysis of student learning data from the University of Hail was conducted for the study [8].The bidirectional long short term model (BiLSTM), was utilized to investigate students whose retention was at risk. The model has diverse features which can be utilized to assess how new students will perform and thus aiding in the timely prediction of student retention and dropout. The method of using Conditional Random Fields (CRF) for sequence labeling was employed to make independent predictions for each label of the students. prediction of student retention was possible with an accuracy of over 0.85 in most scenarios and with FP rates ranging from 0.05 to 0.10 in most cases.

In a study conducted by H. N. Alhulail and H. P. Singh, the primary objective of the research was to develop a model that maximizes predictability and enables the early identification of at-risk student-teachers [9]. The researchers used questionnaires and a four-step logistic regression procedure to analyze a sample of 1723 student-teachers enrolled in public teacher training colleges (TTCs) in a least-developed country (LDC). The study recognized instructional overall performance and aspirations as the most important predictors of student-teacher attrition.

A recent research project aimed to develop a classifier that could predict the academic performance of computer science students at Al-Muthanna University's College of Humanities (MU) [10] The study employed several machine learning techniques, including Naïve Bayes, Logistic Regression, Artificial Neural Network, and Decision Tree, to build predictive models. The models were compared using performance measures such as the ROC index and accuracy, while different metrics such as the F measure, classification error, recall, and precision were computed. To build the models, the researchers used a dataset that combined information from a survey administered to the students and the students' grade book. The ANN model achieved the best performance that is equal to 0.807 and achieved the best accuracy that is equal to 77.04 %.

H. M. R. Hasan, A. S. A. Rabby, M. T. Islam and S. A. Hossain [11] presented a study that aimed students performance prediction using Machine Learning Algorithms.For training they have 1170 students' data of three subjects. They have collected the data from the students record. Finally they were able to get the most perfect result and accuracy with K-Nearest Neighbors, Decision Treee Classifier model with an accuracy of 89.74 % and 94.44 %.

III. METHODOLOGY

A. Dataset Description

The dataset used here is the Student Performance Dataset from a Portuguese secondary school. It contains information on students' personal and academic backgrounds as well as their performance in a subject: Portuguese.

The dataset includes 33 variables, such as the student's age, gender, family size, parents' education, travel time, study time, previous failures, weekly study time, weekly alcohol consumption, whether the student has internet access at home, and their final grades in the course.

The dataset was collected from 2005 to 2006, and it includes 649 instances. The purpose of this dataset is to predict the final grade (G3) of students based on their personal and academic background.

B. Data Preprocessing

Data preprocessing is a crucial step in the data analysis pipeline that involves cleaning, transforming, and preparing raw data for further analysis. The following steps are taken to preprocess the data:

One-hot encode categorical variables using Pandas' get_dummies() function. Encode binary variables using scikit-learn's Label Encoder class. Normalize numeric variables using scikit-learn's Standard Scaler class. Normalization scales the numeric variables so that they have a mean of 0 and a standard deviation of 1. The final preprocessed data consists of the one-hot encoded categorical variables, the label encoded binary variables, and the normalized numeric variables. The preprocessed data is split into training and testing sets using a 80:20 ratio.

Overall, the data preprocessing steps ensure that the data is in a format that is suitable for analysis and prediction using deep learning algorithms. This helps to improve the correctness of the predictions and increases the confidence in the results.

C. Artificial Neural Network(ANN)

ANN model is important in student performance prediction task because it can capture complex non-linear relationships between the input variables and the output variable.

The input layer receives data, which is then passed through one or more hidden layers before reaching the output layer. Each neuron in the network receives inputs from other neurons [12] ,performs a mathematical computation on the inputs, and produces an output that is passed to other neurons in the next layer.

The weights and biases of the connections between neurons are adjusted during training, using an algorithm such as back propagation, to optimize the network's performance on a given task, such as classification, regression, or pattern recognition.

In this study, the main use of ANN is to predict the final grades (G3) of high school students based on various input features such as demographics, academic performance, and family background.

Initially, preprocesses the data by encoding categorical variables, normalizing numeric variables, and splitting the data into training and testing sets. The ANN model is then built using Keras, which consists of an input layer, two hidden layers with dropout regularization, and an output layer. The model is compiled with a mean squared error loss function and an Adam optimizer, and then trained on the training set for 100 epochs with a batch size of 16.

After training, the model is used to make predictions on the test set, which are then compared to the actual test scores. The accuracy of the model is calculated by comparing the predicted binary outcomes (pass/fail) to the actual binary outcomes, and the mean squared error and mean absolute error are also calculated. Finally, a scatter plot is generated to visualize the comparison between the actual and predicted test scores.It obtained an accuracy of 0.846 and the records of MSE is 2.19 and MAE is 1.1.

D. Long Short Term Memory(LSTM)

LSTM works by selectively remembering or forgetting information from previous inputs using specialized units called LSTM cells, which allows it to maintain a long-term memory of past [13] inputs and make accurate predictions based on that memory.

In this study trains an LSTM (Long Short-Term Memory) model on student performance data to predict their final grade in a course. The input data is preprocessed and normalized using one-hot encoding, label encoding, and standard scaling.The input features are reshaped to fit the LSTM input format of (samples, timesteps, features), where each sample corresponds to a single student and each timestep corresponds to a single input feature. The model is built using the Keras API in TensorFlow, which includes one LSTM layer followed by two fully connected layers with dropout to prevent overfitting. The model is then compiled using mean squared error loss and the Adam optimizer. The model is trained on the training set and tested on the testing set. The predictions are converted to binary outcomes indicating whether the student passes or fails the course based on a threshold score of 10. Finally, the accuracy of the model is calculated, and a scatter plot is generated to compare the actual and predicted test scores.It obtained an accuracy of 0.876 and the records of MSE is 1.92 and MAE is 0. 97.

E. Convolutional Neural Network(CNN)

CNNs are useful for analyzing sequential data because they are able to learn patterns and features that are invariant to location.They can detect features regardless of where they occur in the input sequence. This is done through the use of convolutional layers [14], which apply filters to the input sequence to detect patterns and features. MaxPooling layers are also used to downsample the output of the convolutional layers, which helps to reduce the number of parameters and prevent overfitting.

A Convolutional Neural Network (CNN) is used for regression to predict the final grade (G3) of high school students based on various features. CNNs are commonly used for image classification tasks, but they can also be used for time series analysis, which is the case here since the input data is reshaped into a 3D array with dimensions (number of samples, number of features, 1).

In this study, the CNN model consists of two Conv1D layers, each followed by a MaxPooling1D layer, and then two fully connected (Dense) layers with Dropout to prevent overfitting. The model is trained using mean squared error as the loss function and the Adam optimizer. After training, the model is used to predict the final grades of the test set, and the mean squared error and mean absolute error are calculated to evaluate the performance of the model. Finally, a scatter plot is used to compare the actual test scores to the predicted test scores.It obtained an accuracy of 0.853 and the records of MSE is 2.1 and MAE is 1.09.

F. Bi-Directional LSTM

A BiLSTM is a type of Recurrent Neural Network (RNN) that processes the input sequence in both forward and backward directions. This means that the model can learn from both the past and the future context of the input sequence [15] ,which can be helpful for tasks such as sequence prediction and classification.

In this case, the BiLSTM layer is used to process the input features of the student data, which are represented as a sequence of values for each student. The output of the BiLSTM layer is then passed through two additional Dense layers with dropout regularization to predict the final test score.

The use of a BiLSTM layer in this model can help to capture the complex relationships and patterns in the input sequence, and potentially improve the accuracy of the predictions. However, it's important to note that the effectiveness of the BiLSTM layer will depend on the specific task and the characteristics of the input data.The model is compiled with mean squared error loss and the Adam optimizer. It is then trained on the training set for 100 epochs with a batch size of 16. After training, the model predicts final grades of the test set.It obtained an accuracy of 0.823 and the records of MSE is 2.64 and MAE is 1.2.

IV. RESULT

The bar chart displays the accuracy scores of four different algorithms - Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Bidirectional LSTM (BiLSTM) for predicting students performance.

The results show that LSTM achieved the highest accuracy score of 0.876, followed by CNN with an accuracy score of 0.853, ANN with an accuracy score of 0.846, and BiLSTM with an accuracy score of 0.823.

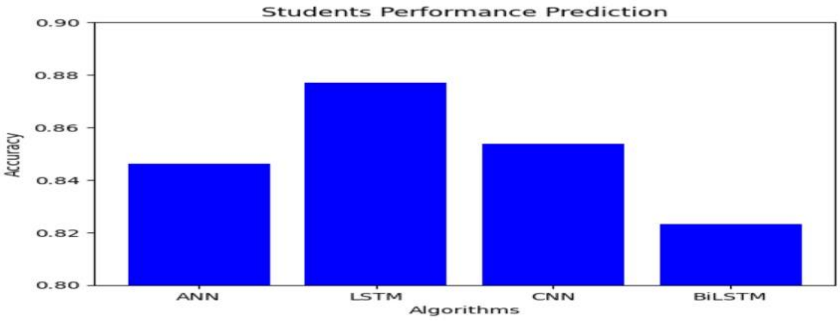

Fig. 1: Bar Chart

LSTM achieved the highest accuracy score, followed by CNN, ANN, and BiLSTM.This may be attributed to the fact that the LSTM algorithm is able to capture long-term dependencies in sequential data, such as student performance records, which may provide more information for accurate prediction.

This helps in visualizing the accuracy scores of different deep learning algorithms for predicting students performance.

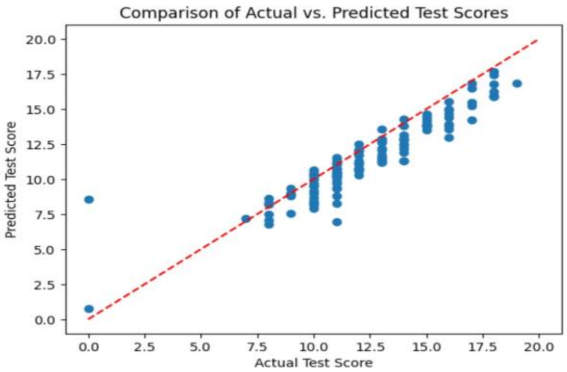

Fig. 2: ANN

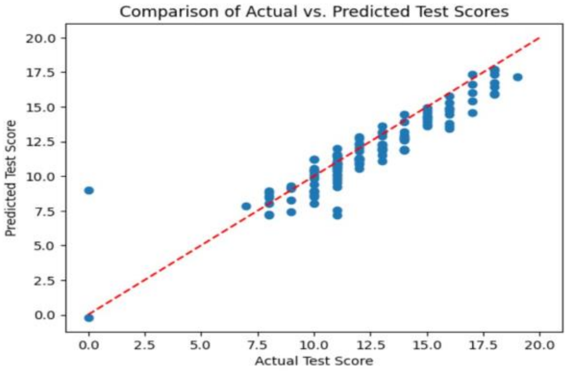

Fig. 3: LSTM

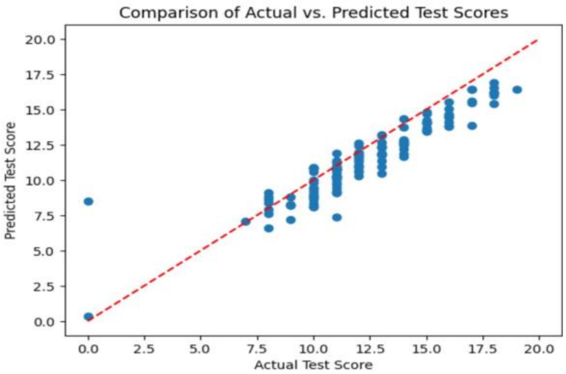

Fig. 4: CNN

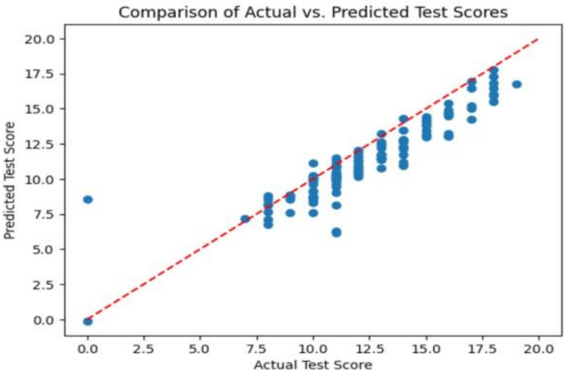

Fig. 5: Bi-LSTM

The scatter plots above compares the actual test scores with the predicted test scores. The diagonal line represents perfect predictions, and the plot helps to visualize the performance of the models in predicting the test scores.

V. CONCLUSION

In the study, we have used Deep Learning model to predict the performance of students using the Portuguese course grades data set.We have used four deep learning algorithms:

ANN,LSTM,CNN,BiLSTM.Based on our project results, the LSTM algorithm achieved the highest accuracy score of 0.876, making it the best-performing algorithm among the four deep learning models used for student performance prediction.

This study can help teachers and school administrators identify students who may need additional support or resources to improve their academic performance. Additionally, the model can provide insights into which factors are most strongly associated with student performance, which can inform educational policies and interventions.

There are several ways in which the above study can be improved. These include collecting more data to improve the model's accuracy and generalization, performing feature engineering to transform the existing features into more meaningful ones, using ensemble learning techniques to improve the model's accuracy and robustness and adding more complex layers to capture more complex patterns in the data.

REFERENCES

[1.] David Ashton, Francis Green, et al. *Education, training and the global economy*. Edward Elgar Cheltenham, 1996.

[2.] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, 40(6):601–618, 2010.

[3.] Aya Nabil, Mohammed Seyam, and Ahmed Abou-Elfetouh. Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, 9:140731– 140746, 2021.

[4.] Bashir Khan Yousafzai, Sher Afzal Khan, Taj Rahman, Inayat Khan, Inam Ullah, Ateeq Ur Rehman, Mohammed Baz, Habib Hamam, and Omar Cheikhrouhou. Studentperformulator: student academic performance using hybrid deep neural network. *Sustainability*, 13(17):9775, 2021.

[5.] Md Rifatul Islam Rifat, Abdullah Al Imran, and ASM Badrudduza. Edunet: a deep neural network approach for predicting cgpa of undergraduate students. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6. IEEE, 2019.

[6.] Abdulaziz Salamah Aljaloud, Diaa Mohammed Uliyan, Adel Alkhalil, Magdy Abd Elrhman, Azizah Fhad Mohammed Alogali, Yaser Mohammed Altameemi, Mohammed Altamimi, and Paul Kwan. A deep learning model to predict student learning outcomes in lms using cnn and lstm. *IEEE Access*, 10:85255–85265, 2022.

[7.] Harjanto Prabowo, Alam Ahmad Hidayat, Tjeng Wawan Cenggoro, Reza Rahutomo, Kartika Purwandari, and Bens Pardamean. Aggregating time series and tabular data in deep learning model for university students' gpa prediction. *IEEE Access*, 9:87370–87377, 2021.

[8.] Diaa Uliyan, Abdulaziz Salamah Aljaloud, Adel Alkhalil, Hanan Salem Al Amer, Magdy Abd Elrhman Abdallah Mohamed, and Azizah Fhad Mohammed Alogali. Deep learning model to predict students retention using blstm and crf. *IEEE Access*, 9:135550–135558, 2021.

[9.] Harman Preet Singh and Hilal Nafil Alhulail. Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach. *IEEE Access*, 10:6470–6482, 2022.

[10.] Hussein Altabrawee, Osama Abdul Jaleel Ali, and Samir Qaisar Ajmi. Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, 27(1):194–205, 2019.

[11.] HM Rafi Hasan, AKM Shahariar Azad Rabby, Mohammad Touhidul Islam, and Syed Akhter Hossain. Machine learning algorithm for student's performance prediction. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2019.

[12.] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[13.] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14.] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[15.] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging.

[16.] *arXiv preprint arXiv:1508.01991*, 2015.