Predictive Monitoring of Learning Processes

Conference Paper · March 2023

DOI: 10.1109/IDCIoT56793.2023.10052784

1 author:

Gomathi Thiyagarajan
CMR Institute of Technology
12 PUBLICATIONS **15** CITATIONS

Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT 2023)
IEEE Xplore Part Number: CFP23CV1-ART; ISBN: 978-1-6654-7451-1

2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 978-1-6654-7451-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/IDCIoT56793.2023.10052784

Predictive Monitoring of Learning Processes

Gomathi Thiyagarajan
Research Scholar (VISTAS, Chennai),
CMR Institute of Technology,
Bengaluru
gomathi.t@cmrit.ac.in

Prasanna S
Professor & Head, ,
School of Computing Science
VISTAS, Chennai
prasanna.scs@velsuniv.ac.in

*Abstract:***What students do in a self-paced online learning environment is a "black box". The instructor has limited interactions with students and a restricted understanding of how students are progressing in their studies. A technology, sophisticated enough to predict the outcome of the student in an online learning environment was widely adopted in Predictive Learning Analytics. In the past, research on predictive learning analytics has emphasized predicting learning outcomes rather than facilitating instructors and students in decision-making or analyzing student behavior. This research study employed a predictive process monitoring technique to analyze the student's event logs in an online learning and online test environment to predict the next activity the student is going to perform and the remaining time to complete the course or test. The Long Short Term Memory neural network approach is used in this work to predict the next activity of the running case by analyzing the sequence of historical data and Apromore to predict the completion time of a case. By employing the predictive monitoring of learning processes, new insights are developed to analyze students' behavior in real-time and is achievable.**

Keywords: Predictive Process Monitoring, Predictive Learning Analytics, Process Mining, Learning Processes

I. INTRODUCTION

In any self-paced learning environment, each student will progress through the curriculum at varying rates and levels. Students who lag often suffer from a negative "ripple effect" wherein the deferred progress leads to discouragement and difficulty mastering complex subjects. The best method to deal with this situation is to provide timely guidance and assistance. Unfortunately, this is frequently easier said than done in real-time. Using event logs, the research field of process mining seeks to identify, track, and enhance real-time processing. Predictive Process Mining (PPM) is a general approach to predict the outcome of the traces or predict the future path of the running trace. PPM often referred to as predictive process monitoring is a technique used to predict the completion time and sequence of activities of a current trace by evaluating the historical data. By comparing specific metrics to what they normally mean for course success, instructors may keep a close check on student development by employing the PPM technique in the learning domain which can also be referred to as Predictive Learning Analytics.Christopher et al. [1] stated that Predictive Learning Analytics (PLA) has the potential to transform instructional techniques and enhance student progress and achievement. However, PLA implementation at academic institutions is still comparatively low, and instructors who do employ PLA don't handle it rigorously. To monitor and manage learners' performance, Mubarak et al. [2] have used the LSTM technique to predict which students are more likely to reduce their interaction with the course videos, hence identifying those who are at risk of performing poorly. Bulut et al. [3] emphasized the necessity of using online formative assessments to develop Learning Analytics models that seem to be predictive and are rooted in theory and instructional design. Despite increased research in the field of learning analytics, Wong et al. [4] have stated that there is no comprehensive predictive model backed by a strong evidence background that educators would use to make effective decisions are witnessed. Predicting the next activities and timestamp of a running case is the current trend in the Process mining domain. Ceci M et al. [5] employed sequential pattern mining to identify partial process models and used extra information about the activities and predicted the next activity and completion time.

Tello Leal et al. [6] developed a prediction model for an IoT domain by defining the phases required to anticipate future activity using the LSTM neural network. To determine the following activity and timestamp, Camargo et al. [7] suggested a method for applying deep learning techniques to train precise models of business process behavior from event logs. In their research work, Pasquadibisceglie et al. [8] demonstrated a novel multi-input, deep learning-based approach for processing multi-view data to achieve predictive accuracy from the varied amount of information in each view. A unique prediction method that applies deep learning with recurrent neural networks and depends heavily on explicit process models was put forth by Joerg Evermann et al. [9]. In a different study, Pasquadibisceglie et al. [10] empirically demonstrated the value of computer vision in

predictive analytics by training 2-dimensional neural networks to forecast the subsequent activity of the running case using the RGB encoding technique. In the current research, we present an LSTM-based prediction model for predicting the sequence of future events from the event logs and remaining time using Apromore. As far as we are aware, no work has been done in the field of learning analytics to predict the next activity and the remaining duration of the learning processes. The remainder of the article is organized in the following manner. The methodology employed in the research is described in Section 2. The observations are covered in Section 3, and the conclusion and future work are presented in Section 4.

II. METHODOLOGY

In this part, we outline the metrics and dataset information utilized to assess the next activity and timestamp of students' across two distinct datasets. We used the Long Short-Term Memory (LSTM) [11] algorithm to predict the next activity of an event log which is represented as a sequence of activities. LSTM is a form of artificial neural network used in Artificial Intelligence technology and deep learning. Unlike conventional feed-forward neural networks, LSTM has feedback connections. The event logs were preprocessed to extract each case's specific activities, which were then used as input for the LSTM Neural network to predict the subsequent activity. The timestamp is predicted by using Aprimore, a process mining platform. We evaluated the performance of prediction on two datasets. Well, known metrics such as Accuracy, Loss, Precision, and Recall were used to evaluate the next activity prediction, and error metrics such as Mean Absolute Error, Root Mean Square Error, normalized mean absolute error, and normalized root means square error were used to evaluate the remaining time prediction.

NAEP Dataset. The eighth-grade students' data from the 2019 NAEP Data Mining Competition are included in this log [12]. There were 410 case ids (student ids) in the dataset used for this study, and 144002 events (actions taken by the student) were recorded against the cases. Table 1 provides the attributes of the NAEP dataset. The event logs were preprocessed to extract all activities and presented in Figure 1. The activities in the event log were converted to acronyms and extracted activities were given as input to the LSTM neural network for activity prediction. The dataset as presented in Table 1 was given as it is to the Aprimore tool for remaining time prediction.

SPLE Dataset. This event log is originated from a self-paced four-week course in programming that was offered in 2021 to undergraduate students at an Indian university (anonymous) through a Learning Management System and an online assessment tool. Both the platforms used captured the activity performed by the students along with the timestamp. Thedata from the assessment tool and LMS were gathered and the minimum features required to apply process mining were processed.Three features namely caseid, activity, and timestamp were extracted for the current study by means of simple python script. The activities from both the dataset were further extracted by using aggregate and group by function.

Table 2 presents the structure of the Self Paced Learning Environment (SPLE) dataset and Figure 2 presents the LSTM input structure.

TABLE 1. NAEP DATASET USED FOR PREDICTION

CaseID	Activity	Resource	Timestamp
2333000289	Enter Item	Directions	16-02-2017 14:40
2333000289	Next	Directions	16-02-2017 14:40
2333000289	Exit Item	Directions	16-02-2017 14:40

TABLE 2. SPLE DATASET USED FOR PREDICTION

Id	Activity	Timestamp
Case001	Pre1	08-02-2021 17:08:00
Case001	Post1	08-13-2021 17:38:00
Case001	Pre2	08-01-2021 19:06:00

```
EI NT EXI EI CC NT EXI EI OC MC MC CC NT
EI NT EXI EI CC NT EXI EI OC MC CB CCAL C
VIS VIS VIS RF LF RF EEB VIS LF EEB RF LF
EI NT EXI EI CC NT EXI EI OC CB CCAL OC M
EE RF RF LF EEB RF LF EEB RF LF EEB RF LF
EI NT EXI EI CC CA CA CA CA CC NT EXI EI
C EC EC EC EC EC CC NT EXI EI EXI EI Yes
EI NT EXI EI CC CA CA CA CA CA CA CA CC C
EI NT EXI EI CC EC EC EC EC NT EXI EI OC
W DW SMOF NT EXI EI RF MK FTC FTC FTC MK
EI NT EXI EI CC NT EXI EI CC NT EXI EI OC
EI NT EXI EI TTS TTS TTS TTS CC NT EXI EI
```

Figure 1. NAEP input for LSTM

```
Pre1 Post1 Pre2 Post2 Pre3 Post3
Pre1 Pre2 Pre3 Pre4 M1H M2H M2H M
Post1 Post2 Post3 Post4 Reg1 Gen1
Pre1 Pre2 Pre3 Pre4 M1H M1H M2H M
Pre1 Post1 Pre2 Post2 Pre3 Post3
Pre1 Pre2 Pre3 Pre4 M1H M2H M3H M
Pre1 Post1 Pre2 Post2 Pre3 Post3
Pre1 Post1 Pre2 Post2 Pre3 Post3
Pre1 Pre2 Pre3 Pre4 FT FT
```

Figure 2. SPLE input for LSTM

III. RESULTS

Keras [13], a Python library for creating models of deep learning networks, was used in the next activity prediction implementation and Apromore is used for time prediction.Table 3 displays the LSTM network's implementation parameters used in the current study for predicting the next activity and the configuration used for predicting the time.The LSTM neural network was trained for two different event logs namely NAEP and SPLE as described in the dataset section.The event log NAEP includes 410 cases and 42 activities, and SPLE includes 549 cases and 25 activities.

The LSTM neural network during its training identified the vocabulary size of 46 and 26, and the number of sequences as 144002 and 7681 for the NAEP dataset and SPLE dataset respectively.The event log was preprocessed to extract only the activities attribute from the log and given as input to the LSTM network to predict the next activity. Each instance of the neural network was set up to predict three outputs, ranked from highest to lowest probability.

Both the dataset used in the study as the individual set is split into 20% of the validation set and 80% of the training set. The data set is fed to the layered neural networks. An output layer is built after the hidden layer (LSTM) and input layer (embedding). Table 4 below lists the output shape, layer type, and Param#.Figures 3 and 4 show the loss, accuracy, precision, and recall data of the model for each epoch, while Table 5 gives an excerpt of the results from the neural network's prediction of the next activity.

TABLE 3. PARAMETER FOR THE PREDICTION

Parameter	Value
Next Activity	
Epochs	50
Batch Size	20
LSTM Units	50
Activation Function	Softmax
Loss	Categorical_crossentrophy
Optimizer	Adam
Metrics	Accuracy, Precision, Recall
Remaining time	
Prediction Type	Remaining time
Prediction Method	Cat boost
Feature Encoding	Aggregate
Metrics	MAE, NMAE, RMSE, NRMSE

TABLE 4. LSTM LAYERS, OUTPUT SHAPE, AND PARAM

Dataset	Layer	Output Shape	Param#
NAEP	Embedding	<None,2,50>	1300
	LSTM	<None,50>	20200
	Dropout	<None,50>	0
	Dense	<None,26>	1326
SPLE	Embedding	<None,2,50>	2300
	LSTM	<None,50>	20200
	Dropout	<None,50>	0
	Dense	<None,46>	2346

The Matplotlib library was used to generate these graphs. A loss curve during training is one of the most frequently used plots for debugging a neural network. It gives a preview of the training process and the direction in which the network learns.The epoch vs loss graph in Figure 3 and Figure 4 shows a good learning rate. To understand the progress of neural network accuracy vs epoch curve was plotted. The gap between the training and the testing accuracy is a clear indication of overfitting. The gap is minimal in the SPLE dataset compared to NAEP. Apromore [14] combines a variety of process-specific feature engineering and prediction bucketing techniques with machine learning algorithms to produce estimates that are both accurate and reliable. The objective of such predictions is to provide the decision-makers with accurate, stable predictions that can be presented to them as soon as possible to shorten the time it takes for critical intervention. Using the monitor option available in the Apromore tool the remaining time of each case was predicted. The prediction was made using the aggregate encoding technique and the cat boost prediction method. Using accuracy vs. prefix length and finished time for both datasets, the mean absolute error, root mean square error, normalized mean absolute error, and normalized root mean square error was calculated and shown in Table 6. Table 7 and Table 8 presents the output generated for predicting the time for the NAEP and SPLE dataset. Visualization of accuracy and prefix length is given in Figure 5 and Figure 6.

TABLE 5. RESULTS FROM LSTM NEURAL NETWORK

Dataset	Accuracy	Loss	Precision	Recall
NAEP	0.7133	0.8870	0.7988	0.6407
SPLE	0.7673	0.6316	0.7971	0.7148

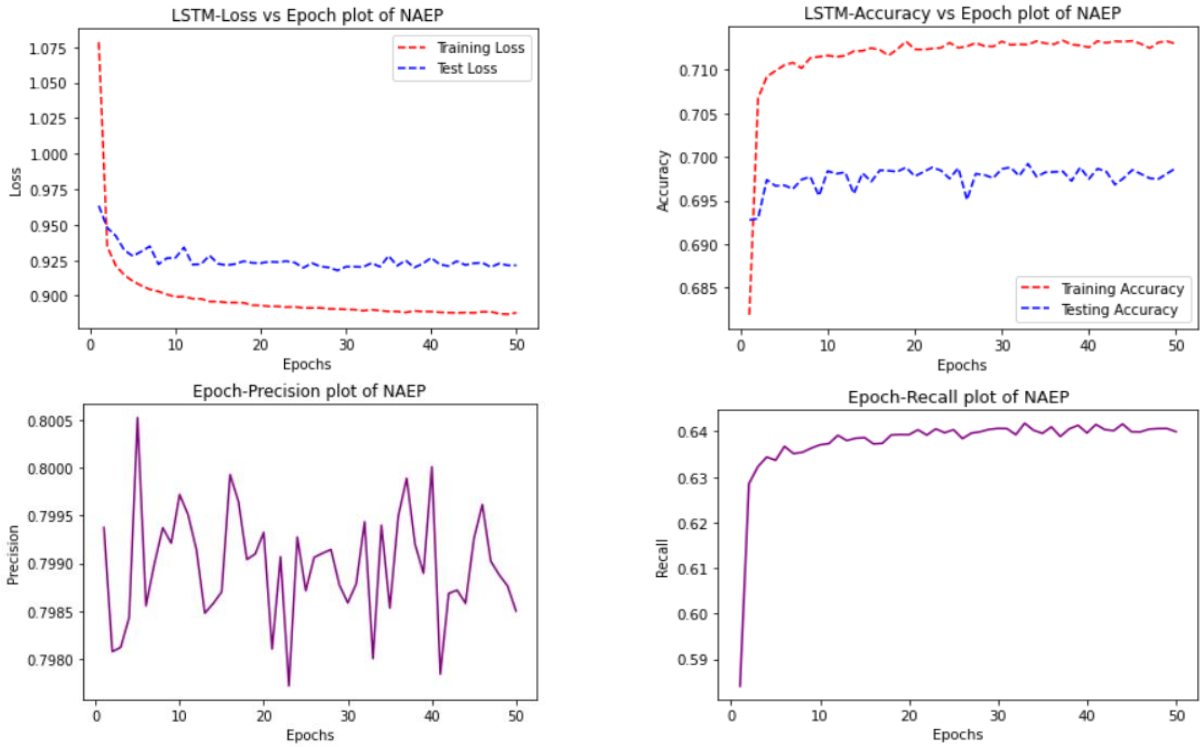

Figure 3. Visualization of Epoch vs loss, accuracy, precision, and recall for the NAEP dataset

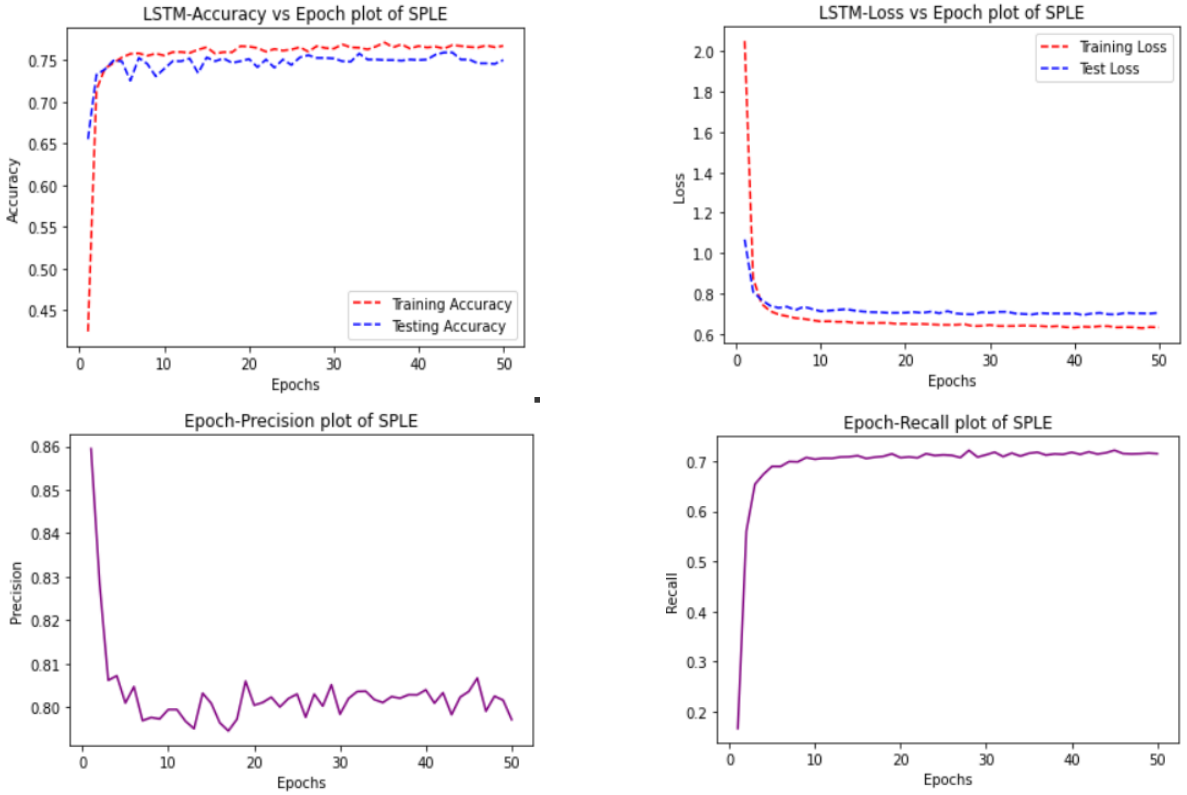

Figure 4. Visualization of Epoch vs loss, accuracy, precision, and recall for the SPLE dataset

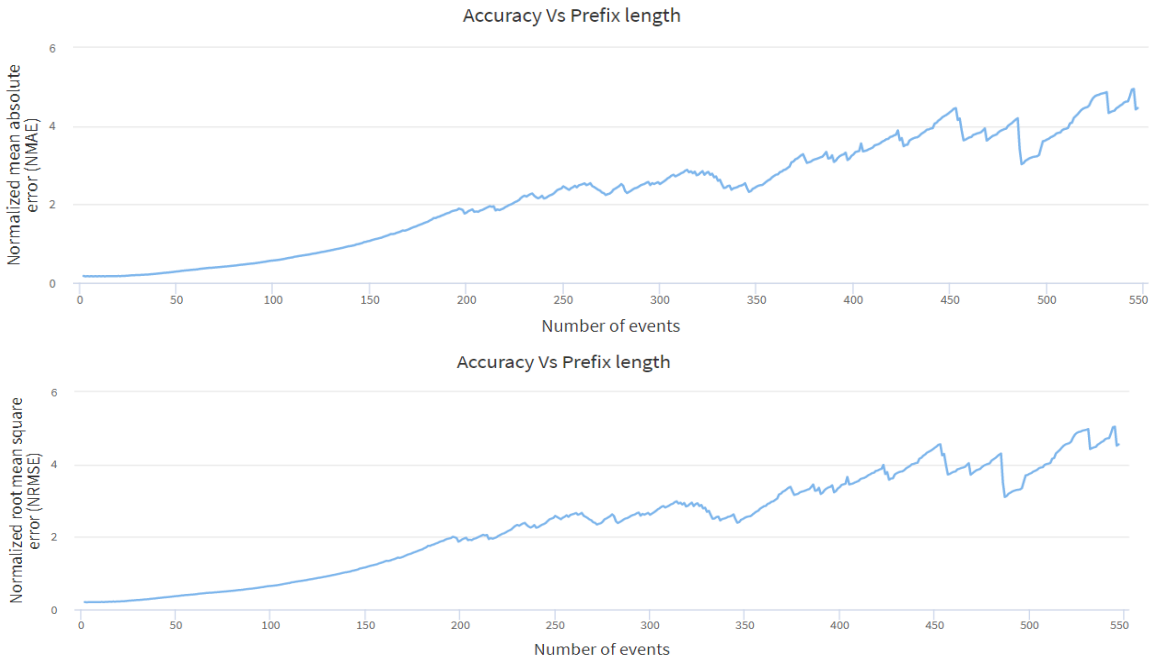

Figure 5. Visualization of Accuracy vs Prefix length for NAEP dataset

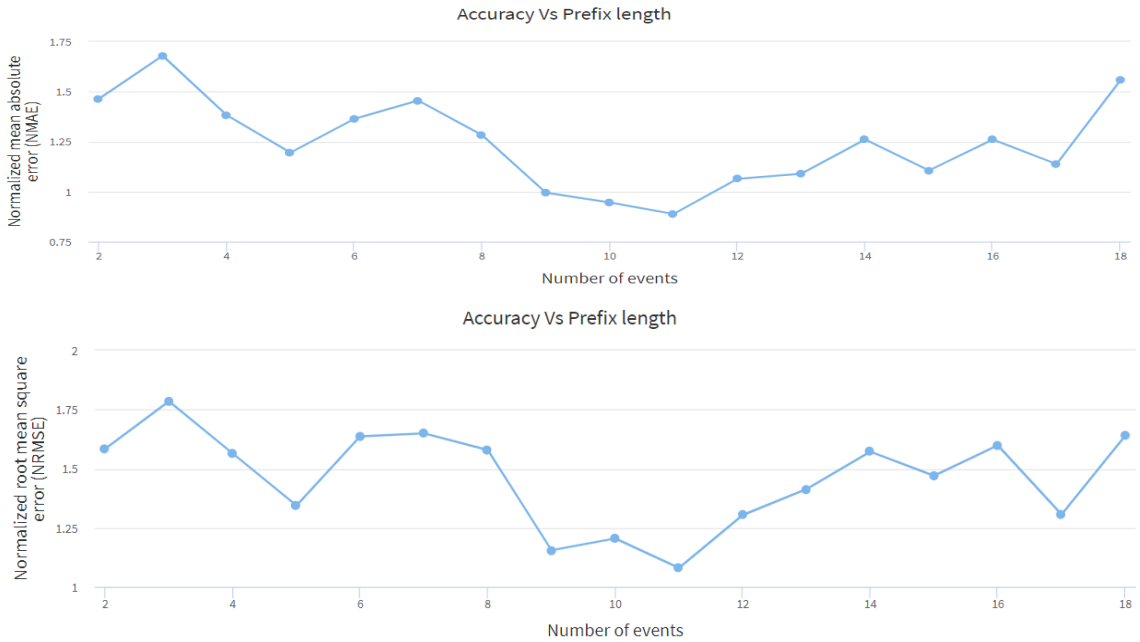

Figure 6. Visualization of Accuracy vs Prefix length for SPLE dataset

TABLE 6. RESULTS FROM APROMORE FOR TIME PREDICTION

Dataset	MAE	NMAE	RMSE	NRMSE
NAEP	757.37	1.02	870.13	1.17
SPLE	6748892.4	1.34	7806134.62	1.55

TABLE 7. PREDICTION OF REMTIME AND PREDICTED COMPLETION FOR NAEP DATASET

Case_id	Remtime	Last_timestamp	Predicted Completion
Case001	46598.31	08-03-2021 08:12	08-03-2021 21:08
Case002	90775.32	08-03-2021 07:55	09-03-2021 09:07
Case003	137551.9	08-03-2021 07:59	09-03-2021 22:11

TABLE 8. PREDICTION OF REMTIME AND PREDICTED COMPLETION FOR SPLE DATASET

Case_id	Remtime	Last_timestamp	Predicted Completion
2333000289	1459.415996	16-02-2017 04:08	16-02-2017 04:32
2333001633	1566.228698	02-02-2017 03:28	02-02-2017 03:54
2333001897	1455.267	01-03-2017 07:47	01-03-2017 08:11

IV.CONCLUSION AND FUTURE WORK

In this research, we have used Neural Networks with process mining to predict the next activity and remaining time of a case in an event log. We empirically showed the performance of the techniques on the two datasets NAEP and SPLE and presented it in the results section. The next-activity prediction that can be made using Predictive Process Mining in process executions is calculated by identifying the probabilities of each possible scenario which focuses on ongoing cases and predict three outputs. The LSTM neural network was trained with specific parameters to meet the objective. The parameters of the model include 50 epochs, with batch size set to 20 and LSTM units set to 50. Softmax activation function was used for training the LSTM network. The loss was calculated by categorical_crossentrophy and Adam was the optimizer employed. The result shows the accuracy of NAEP and SPLE datasets for the next activity prediction was 0.71 and 0.76 respectively. Apromore a popular web-based tool was used to predict the remaining time with aggregate encoding and the cat boost prediction method. The NMAE and NRMSE were recorded as 1.02 and 1.17 for the NAEP dataset and 1.34 and 1.55 for the SPLE dataset. With the help of the predictive process mining that was used in the learning processes, we can get useful insights about what should be done next based on event logs. The use of event logs from the education domain is what makes this implementation unique. The validation of the current approach employed in learning processes required additional tests and it was also observed the order in which the traces are used to train the neural network has a direct impact on the results due to LSTM cell states and hence results obtained cannot be generalized to other cases.

We believe the ability to analyze event logs from a learning or test environment in real-time can provide valuable insights about student behavior and support by providing timely inputs for successful learning outcomes. We would like to extend the research by comparing the current approach employed with other state of art techniques to predict the next event and remaining time. Another potential for future work is to predict all events of a case to completion.

REFERENCES

[1] ChristotheaHerodotou, Claire Maguire, Nicola McDowell, Martin Hlosta, AvinashBoroowa, “The engagement of university teachers with predictive learning analytics”, Computers & Education, Volume 173, 2021.

[2] Mubarak, A.A., Cao, H. & Ahmed, S.A,”Predictive learning analytics using deep learning model in MOOCs’ courses videos”.,Educ Inf Technol 26,2021, pp 371–392

[3] Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., Lai, K. W., & Shin, J,” Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models”, British Journal of Educational Technology, 2022,pp1– 21.

[4] Wong, B.Tm., Li, K.C,” A review of learning analytics intervention in higher education (2011–2018)”, J. Comput. Educ. 7, 2020, pp 7–28.

[5] Ceci, M., Lanotte, P.F., Fumarola, F., Cavallo, D.P., Malerba, D, “Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining”, In Dzeroski, S., Panov, P., Kocev, D., Todorovski, L. (eds) Discovery Science. DS 2014. Lecture Notes in Computer Science, vol 8777. Springer, Cham., 2014.

[6] E. Tello-Leal, J. Roa, M. Rubiolo and U. M. Ramirez-Alcocer, "Predicting Activities in Business Processes with LSTM Recurrent Neural Networks”,Machine Learning for a 5G Future (ITU K), 2018, pp. 1-7.

[7] Camargo, M., Dumas, M., & Rojas, O.G” Learning Accurate LSTM Models of Business Processes”, International Conference on Business Process Management,2019.

[8] V. Pasquadibisceglie, A. Appice, G. Castellano, and D. Malerba, "A Multi-View Deep Learning Approach for Predictive Business Process Monitoring," in IEEE Transactions on Services Computing, vol. 15, no. 4, pp. 2382-2395, 2022.

[9] Joerg Evermann, Jana-Rebecca Rehse, Peter Fettke, “Predicting process behavior using deep learning, Decision Support Systems”, Volume 100,2017, Pages 129-140, ISSN 0167-9236.

[10] Pasquadibisceglie, V., Appice, A., Castellano, G., Malerba, D,”Predictive Process Mining Meets Computer Vision. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds) Business Process Management Forum. BPM 2020. Lecture Notes in Business Information Processing, vol 392. Springer, Cham.2020.

[11] “Deep Learning Introduction to Long Short Term Memory” retrieved from https://www.geeksforgeeks.org on 2nd November 2022.

[12] NAEP Data Mining Competition for 2019. https://sites.google.com/view/dataminingcompetition2019/dataset?authuser=0

[13] “Deep Learning for Humans” retrieved from https://keras.io/ on 2nd November 2022.

[14] “Predictive Process Monitoring” retrieved from https://apromore.com on 15th November 2022