Advancing Early Intervention Strategies: Leveraging Shorter xAPI Data Sequences with LSTM for At-Risk Student Detection

James Bilitski, Ph.D.
University of Pittsburgh
Johnstown, PA
bilitski@pitt.edu

Paul Jesukiewicz
Powertrain, Inc.
Hyattsville, MD
PJesukiewicz@powertraiun.com

Jonathan Poltrack
Veracity Technology Consultants, LLC
Johnstown, PA
jono@veracity.it

ABSTRACT

As the landscape of education increasingly incorporates online and technology-assisted learning environments, managing large student populations has become a critical challenge. This study focuses on enhancing early intervention strategies to improve student outcomes by leveraging shorter Experience API (xAPI) data sequences with Long Short-Term Memory (LSTM) networks for at-risk student detection. The primary objectives are to establish a model for measuring student risk and apply machine learning techniques to identify at-risk students based on their activity patterns, ultimately facilitating timely interventions.

The dataset used in this study, derived from a large defense agency, includes over 18 million xAPI statements capturing a wide range of learner interactions within a Learning Management System (LMS). The study utilized an LSTM network with eight hidden layers, each containing 50 units, designed to process shorter sequences of learner interactions to predict outcomes.

Various sequence lengths (10, 20, 30, 50, 60, 75, and 100) were tested to determine the optimal sequence length for early detection. The results showed that validation accuracy generally improved with increasing sequence length, plateauing at a sequence length of 75, where the model achieved a validation accuracy of 78.1%. This indicates that shorter sequences can be effectively used for early prediction, enabling interventions well before critical assessments.

To address class imbalance in the dataset, an oversampling technique was employed, balancing the number of passes and fails used for training. This approach ensured that the model was trained on a representative dataset, enhancing the robustness and accuracy of predictions.

The study's findings highlight the potential of using deeper LSTM networks to capture complex patterns in sequential data, thereby improving early detection of at-risk students. The ability to make accurate predictions with shorter sequences allows for earlier interventions, providing timely support to at-risk students and ultimately improving educational outcomes. Integrating these predictive models into real-world educational systems will provide valuable insights and support for at-risk students, enhancing their chances of success.

ABOUT THE AUTHORS

James Bilitski, Ph.D. is an associate professor of computer science at the University of Pittsburgh at Johnstown and an industry consultant. He has worked for companies such as Motorola, Phillips, Ansaldo, Concurrent Technologies, Clair Global, and Problem Solutions. Dr. Bilitski has led technical projects in commercial and government spaces. He specializes in machine learning, artificial intelligence, education, real time systems, and audio and music software. He has several peer reviewed publications in the areas of machine learning.

Paul Jesukiewicz is a leader in the field of learning technologies with over 35 years of experience working in government, industry, and academia. He successfully led research, development, and implementation of a global program on advanced distributed learning. He was inducted into the Federal Government Distance Learning Association (FGDLA) Hall of Fame in 2012 as recognition for his significant career accomplishments in promoting and developing distance learning in the Federal Government. He is currently the Vice President of Products at Powertrain.

Jonathan Poltrack has worked in the learning industry for over 20 years, with large periods of time at the DoD's ADL Initiative. At the ADL Initiative, Jonathan was an early contributor to the Sharable Content Object Reference Model (SCORM), which became a de facto global e-learning specification. Later at ADL, Jonathan began leading efforts aimed at transitioning SCORM while specifying a new learning platform based on modern technologies and software architectures including the xAPI. Jonathan co-founded Veracity Technology Consultants, a company that focuses on standards-based learning technology services and products, with several learning technology expert partners. Jonathan is passionate about education, training, and performance support and their intersections with technology.

Advancing Early Intervention Strategies: Leveraging Shorter xAPI Data Sequences with LSTM for At-Risk Student Detection

James Bilitski, Ph.D.
University of Pittsburgh
Johnstown, PA
bilitski@pitt.edu

Paul Jesukiewicz
Powertrain, Inc.
Hyattsville, MD
PJesukiewicz@powertraiun.com

Jonathan Poltrack
Veracity Technology Consultants, LLC
Johnstown, PA
jono@veracity.it

INTRODUCTION

As the field of education increasingly embraces online and technology-assisted learning environments, managing large student populations effectively becomes paramount. The diversity of content and the variety of hardware devices used in these environments further complicate the management process. Industry and government systems often include hundreds of e-learning courses without a dedicated facilitator or instructor, which underscores the need for advanced technology-assisted tools to manage students' progression through curricula in online learning platforms.

The primary objective of this project is twofold: first, to establish a model for measuring the risk of students, and second, to apply machine learning (ML) techniques to identify at-risk students by discovering activity patterns that lead to passing or failing assessments. Early detection mechanisms are crucial as they facilitate timely interventions, increasing the likelihood of student success (Pek et al., 2023). The culmination of this project will be a suite of tools within the Government Learning Enclave (GLE). These tools will enable instructors, mentors, and administrators to retrieve lists of potential at-risk students, gauge their likelihood of failing assessments, and understand the factors contributing to their at-risk status. This focused approach will better equip administrative users to assist the most vulnerable students, leveraging insights provided by ML.

Government Learning Enclave (GLE)

The FedRAMP authorized Government Learning Enclave is an ideal environment for researching the identification and intervention of at-risk students. GLE provides learning products and services to hundreds of U.S. government clients, serving over ten million end users. Many GLE systems are standards-based, ensuring consistent and broadly adopted data formats. The ecosystem includes Learning and Talent Management Systems (LMS/TMS), Learning Record Stores (LRS), Learning Content Management Systems (LCMS), and Student Information Systems (SIS). The majority of the training provided is professional, mandatory online training for government agencies.

The GLE became FedRAMP authorized in 2017 and subsequently accumulated large amounts of learners' experiential data in its LRS instances. xAPI data offers granular insights into learner interactions with content, such as course completions, assessment responses, performance evaluations by instructors, and interactions with virtual reality (VR) experiences. The extensive tracking capabilities of xAPI, which surpass previous learning standards, made it an ideal choice for this project. The availability of significant amounts of xAPI data, representing a wide range of training content and interactions, further reinforced this choice.

The dataset used in this project is a real learning dataset from the GLE collected from September 1, 2019, to August 31, 2022. This dataset, comprising over 18 million xAPI statements, was preprocessed before use. It captures learner interactions across online courses, virtual instructor-led training (VILT) courses, and online assessments. Although specific course titles cannot be disclosed, the subjects covered a broad spectrum of government professional training

topics, including job-specific instruction, information systems, cybersecurity, trafficking in persons awareness, ethics, and more. The assessments included various types of content, such as inline knowledge checks, pre-tests, and post-tests.

This study builds on the foundation of previous research and leverages the extensive data available within the GLE ecosystem. By developing a deeper understanding of student interactions and applying sophisticated ML models, this project aims to improve early detection and intervention strategies for at-risk students, ultimately enhancing their chances of success.

LITERATURE REVIEW

The landscape of higher education has increasingly embraced online learning environments, offering flexibility and accessibility to a diverse student population. However, this flexibility, particularly in self-paced learning programs, introduces significant challenges in student retention and success. Predicting at-risk students early in their educational journey is required for timely interventions that can improve outcomes. This literature review explores existing research on machine learning techniques, particularly Long Short-Term Memory (LSTM) networks, for predicting at-risk students. It also highlights the gaps addressed by the present study, which leverages shorter Experience API (xAPI) data sequences for early detection.

Self-paced learning environments are designed to cater to individual student needs, allowing them to progress through course material at their own pace. However, this approach often leads to increased risks of student disengagement and dropout due to the lack of structured deadlines and the need for strong self-regulation skills. Waheed et al. (2023) emphasized the necessity of early prediction in self-paced learning, noting that early identification of at-risk students can enable timely interventions and support structures.

Machine learning (ML) has been used for prediction in a wide variety of fields, including the prediction of student risk, exam scores, and early identification of unsuccessful students. The identification of at-risk students and providing intervention has gained significant attention among the research community (Pek et al., 2023; Al Breiki et al., 2019; Chui et al., 2019; Er et al., 2012; Jang, 2022; Lakkaraju, 2015; Livieris, 2018; Macarini, 2019; Pilotti et al., 2022). Various ML prediction algorithms have been employed, such as Deep Networks, Decision Trees, Random Forests, Regressions, Naïve Bayes, and Support Vector Machines. ML techniques are important in extracting information and knowledge from datasets. For instance, Adnan (2021) developed a system to predict at-risk students by analyzing performance during different modules in a course. Pek (2022) conducted a study to determine if initial student performance is informative in the early detection of at-risk students.

LSTM networks, a type of recurrent neural network, are particularly well-suited for tasks involving sequential data due to their ability to retain information over long periods. Waheed et al. (2023) employed LSTM networks to predict at-risk students in a self-paced learning environment, achieving high predictive accuracy by analyzing week-wise aggregated data. Their study highlighted the potential of LSTMs in educational settings but also indicated the need for more granular, earlier predictions.

Bilitski et al. (2023) conducted a study titled "A Machine Learning Approach for Identifying At-Risk Students in Learning Record Stores: A Case Study Using USALearning Experience API (xAPI)," which developed a model to measure student risk by analyzing xAPI data. This study used a sequence size of 100 events to predict student outcomes immediately before an assessment, providing valuable insights but limiting the window for timely interventions. The current study builds on this work by exploring the use of shorter sequences to enable earlier detection of at-risk students.

Self-paced learning programs face unique challenges, including student isolation, lack of social integration, and difficulties in time management. These factors contribute to higher dropout rates compared to more structured learning environments. Identifying at-risk students early in the course can mitigate these issues by providing

targeted support. However, the existing literature often focuses on aggregated data over extended periods, which delays the identification of at-risk students until later in the course.

The present study builds on previous research by investigating the use of shorter xAPI data sequences for earlier predictions. By testing sequence lengths of 10, 20, 30, 50, 60, 75, and 100, this study aims to provide earlier warnings of student risk, enabling interventions well before critical assessments. This approach addresses the gap in the literature related to the timing of predictions and the granularity of data sequences used.

Class imbalance is a common issue in educational datasets, where instances of at-risk students are often outnumbered by those of successful students. Effective techniques for managing class imbalance, such as clustering for downsampling, have been employed to improve the robustness of predictive models. Waheed et al. (2023) used such techniques to ensure their model accurately identified at-risk students despite the imbalance.

Identifying key predictors of student success is important for developing effective intervention strategies. Features such as quiz participation, course homepage interaction, and assessment submission have been found to significantly impact student performance. The current study aims to extract specific attributes from shorter data sequences that drive prediction decisions, thereby providing actionable insights for educators.

In conclusion, the literature emphasizes the importance of early prediction in self-paced learning environments and the effectiveness of LSTM networks in educational data mining. By leveraging shorter xAPI data sequences, the present study seeks to enhance the timeliness and accuracy of at-risk student detection, offering a novel approach to early intervention strategies. This research builds on previous work and contributes to the ongoing effort to improve educational outcomes through data-driven insights and timely support for at-risk students.

STUDENT DATA

The dataset used in this study is derived from the Government Learning Enclave (GLE) Learning Record Store (LRS) built on Experience API (xAPI), which tracks a wide range of student interactions within a Learning Management System (LMS). The data collection spans from September 1, 2019, to August 31, 2022, covering a comprehensive period of student activity. This dataset includes:

- Over 18 million xAPI statements
- Approximately 65GB
- Timeframe: September 1, 2019, to August 31, 2022
- 130,561 students

The xAPI format is a standardized data interchange format that utilizes JavaScript Object Notation (JSON) to represent each event or interaction. To facilitate analysis, the xAPI JSON data was converted into flat comma-separated value (CSV) files, which are compatible with various data science and ML libraries. In xAPI, each statement captures a specific interaction between a student and the learning content. A typical xAPI statement includes:

- Actor: Typically the student who performed the action
- Verb: The action taken by the student
- Object: The activity or content the student interacted with
- The verbs in xAPI statements provide essential context for understanding student interactions. For example, a "scored" verb indicates that the student completed an activity resulting in a score, such as finishing a quiz.

The dataset includes various xAPI verbs representing different types of student interactions with the LMS. These verbs include:

- Viewed: The student viewed a course, assessment (quiz), or other resource or activity in the system.
- Responded: The student answered a question on an assessment or feedback survey.

- Completed: The student completed a course, module, Sharable Content Object (SCO), or other content.
- Initialized: The student launched a SCO in a Sharable Content Object Reference Model (SCORM) package.
- Logged into: The student logged in to the LMS.
- Scored: The student received a score on an assessment, quiz, SCORM package, or other scored content.
- Terminated: The student completed an attempt on a SCO in a SCORM package.
- Passed: The student passed the objective of a SCORM package, course, assessment, or other content.
- Received: The student received a grade from an instructor.
- Failed: The student received a failing score on a SCORM package or assessment.

A sequence in this context refers to an ordered series of xAPI verbs for a student, capturing their interactions with the LMS over time. Each sequence provides a chronological record of a student's actions, enabling the analysis of their learning behaviors and performance patterns. For example, a sequence might include the following xAPI verbs: "logged into," "viewed," "responded," "completed," and "scored." This sequence indicates that the student logged into the LMS, viewed a course, responded to assessment questions, completed the course, and received a score. By analyzing these sequences, the model can identify patterns that are indicative of a student's likelihood to pass or fail, thus enabling early detection of at-risk students.

IMPROVED MODEL FOR PREDICTING STUDENT OUTCOMES

Introduction to LSTM Networks

Long Short-Term Memory (LSTM) networks are a specialized type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. Unlike traditional RNNs, LSTMs effectively manage long-term dependencies through their unique architecture, which includes memory cells and gating mechanisms. These gates regulate the flow of information, allowing LSTMs to retain or discard information as needed. This capability makes LSTMs particularly well-suited for tasks involving time series or sequential data, such as predicting student outcomes based on their interactions over time.

Original LSTM Model

In the previous study, "A Machine Learning Approach for Identifying At-Risk Students in Learning Record Stores: A Case Study Using USALearning Experience API (xAPI)" (Bilitski et al., 2023), an LSTM network was developed to predict student outcomes. The original model used a sequence length of 100 and included fewer hidden layers. The mathematical formulation of the LSTM model is as follows:

The Forget Gate decides what information should be thrown away from the cell state. It uses a sigmoid activation function as shown in equation 1:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

where W_f is the weight matrix for the forget gate, b_f is the bias, h_{t-1} is the previous hidden state, and x_t is the current input.

The Input Gate updates the cell state with new information. It uses two equations (2 and 3):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

where W_i and b_i are the weight and bias for the input gate, and W_C and b_C are the weight and bias for creating a vector of new candidate values, $\tilde{C}_t$, for the state.

The Cell State C_t is updated using the forget gate and the input gate (shown in equation 4):

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (4)$$

where C_{t-1} is the previous cell state.

The Output Gate decides the next hidden state h_t, using the updated cell state and the input shown in equation 5 and 6:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o \quad (5)$$
$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

where W_o and b_o are the weight and bias for the output gate. The hidden state h_t can be used for predictions, while the cell state C_t helps the LSTM to keep track of the underlying contextual information in the sequence.

Improved Model for Predicting Student Outcomes

Building on the foundation of the original model, the improved LSTM network incorporates a deeper architecture with more layers. The new model consists of eight hidden layers, each with 50 units. This enhanced structure, featuring eight hidden layers of 50 cells each, was identified as promising through preliminary research focused on predicting student outcomes using shorter sequences.

Adding more layers to an LSTM network can significantly enhance its ability to learn complex patterns and dependencies in sequential data. The additional layers enable the model to capture hierarchical representations of the data, which can improve its generalization capabilities. This hierarchical learning is particularly beneficial when dealing with shorter sequences, as it allows the model to extract meaningful features from limited data. However, increasing the number of layers also has drawbacks. Deeper networks require more computational resources and longer training times. They are also more prone to overfitting, especially if the dataset is not large enough or if appropriate regularization techniques are not applied.

The size of each layer, or the number of units per layer, also plays an important role in the model's performance. Larger layers can capture more information and learn more complex patterns, providing greater flexibility to fit the training data. This increased capacity can be particularly advantageous when working with shorter sequences, as it allows the model to make accurate predictions with limited data. However, larger layers also increase the risk of overfitting and require more memory and computational power. Balancing the layer size with the overall complexity of the network is essential to achieve optimal performance.

In the context of this study, the raw data comprised a pass count of 85,858 and a fail count of 40,570. To address this class imbalance, an oversampling technique was employed to even out the number of passes and fails used for training. Specifically, the fail count was oversampled to match the pass count, resulting in a balanced dataset with 85,858 passes and 85,858 fails. The oversampling process involved first separating the sequences into two groups based on their labels: passed and failed. The number of failed sequences needed to balance the dataset was then calculated. Random sampling with replacement was used to select additional fail sequences until their count matched the number of pass sequences. This random selection ensured that the oversampling process did not introduce any systematic bias into the training data.

Once the sequences were balanced, they were combined and shuffled to create a mixed dataset for training. This approach allowed the model to learn equally from both pass and fail sequences, improving its ability to generalize across different types of student interactions. By using this method, the study ensured that the model was trained on a representative and balanced dataset, enhancing the robustness and accuracy of its predictions.

The new LSTM model with eight hidden layers of 50 units each is designed to improve the ability to predict student outcomes using shorter sequences. The additional layers enhance the model's capacity to learn from the sequential data, capturing more subtle and complex patterns that might be missed by a shallower network. This deeper

architecture enables the model to make accurate predictions earlier in the learning process, providing timely warnings about at-risk students. By utilizing the increased depth of the network, the improved model can effectively handle the complexities of student interactions within the LMS, leading to more reliable and actionable predictions.

RESULTS

The primary objective of this study was to enhance the predictive capabilities of an LSTM model for identifying at-risk students by leveraging shorter xAPI data sequences. To this end, various sequence lengths were tested, ranging from 10 to 100, to determine the optimal sequence length for early detection. The model's performance was evaluated using metrics such as loss, accuracy, validation loss, and validation accuracy.

Table 1 below presents the performance metrics for different sequence lengths, showcasing the loss, accuracy, validation loss, and validation accuracy for each configuration.

Table 1: Results with Shortened Sequences

Sequence Length	Loss	Accuracy	Validation Loss	Validation Accuracy
10	62.2	61.9	62.8	61.5
20	57.2	65.2	59.1	64.4
30	51.7	70.7	54.0	69.5
50	44.7	75.7	47.4	73.9
60	39.6	79.5	43.7	77.0
75	39.1	80.2	42.4	78.1
100	43.4	76.7	44.9	78.1

The performance metrics indicate that as the sequence length increases, the model's validation accuracy generally improves. This trend continues until a sequence length of 75, after which the accuracy plateaus. Figure 1 illustrates this plateau.

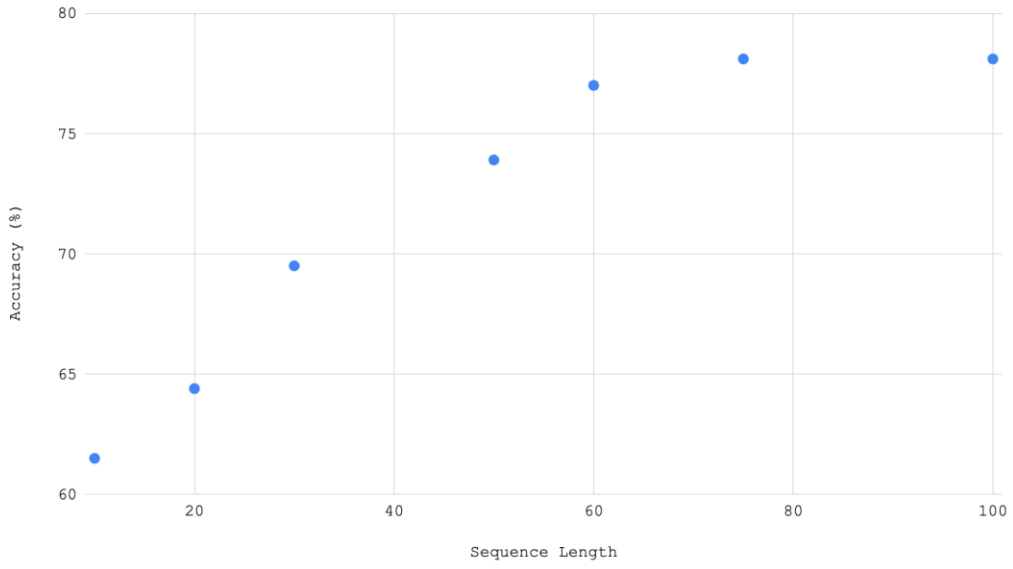

Figure 1: Accuracy with Varying Sequence Lengths

Analysis of Results

The validation accuracy for different sequence lengths is graphically represented in the figure below, illustrating the relationship between sequence length and model performance.

The results show that with a sequence length of 10, the model achieves a validation accuracy of 61.5%, which is relatively low. Increasing the sequence length to 20 results in a validation accuracy of 64.4%. With a sequence length of 30, the model's validation accuracy improves significantly to 69.5%. Further increasing the sequence length to 50 yields a validation accuracy of 73.9%. At a sequence length of 60, the model reaches a validation accuracy of 77.0%. The highest validation accuracy of 78.1% is achieved with sequence lengths of both 75 and 100, indicating no further improvement beyond this point.

DISCUSSION

The enhanced LSTM model, featuring eight hidden layers with 50 units each, demonstrates significant promise in predicting at-risk students using shorter sequences. The model's ability to achieve high validation accuracy with sequence lengths as short as 75 and 60 underscores its effectiveness in early detection, which is needed for timely interventions.

Adding more layers to the LSTM network enhances its feature extraction capability, allowing the model to capture more complex patterns and dependencies within the sequential data. This hierarchical learning enables the model to derive more detailed representations, improving its generalization capabilities. However, increasing the number of layers also introduces certain drawbacks. The deeper network architecture requires more computational resources and longer training times. Moreover, the risk of overfitting is higher, particularly if the dataset is not large enough or if adequate regularization techniques are not applied.

The size of each layer, or the number of units per layer, also plays a role in the model's performance. Larger layers provide greater capacity to learn from the data, which is advantageous when working with shorter sequences. This increased capacity allows the model to make accurate predictions with limited data. However, larger layers also necessitate more memory and computational power, and they can increase the risk of overfitting.

The results of this study suggest that the new LSTM model architecture, with its deeper network structure, can effectively handle the complexities of student interactions within the LMS. By making accurate predictions with shorter sequences, the model allows for earlier identification of at-risk students. This capability enables educators to provide timely support and interventions, ultimately improving student outcomes. The study demonstrates that leveraging shorter xAPI data sequences with an enhanced LSTM model offers a viable approach for early detection of at-risk students, facilitating proactive and targeted educational interventions.

The LSTM model shows great promise for predicting student outcomes. The model converts a probability value into a binary category of pass or fail by comparing it to a threshold value of 0.5. Any value over 0.5 is classified as a pass, while values below 0.5 are classified as a fail. For instance, a student prediction of 0.85 would be classified as a pass, and a student prediction of 0.52 would also be classified as a pass. Binary classification systems inherently have errors for prediction values that fall in the midpoint between categories. While these middle-sitting values generally cause errors in prediction systems, the raw prediction should provide insight when making decisions about intervention. Middle prediction values suggest that a student might fail, warranting an appropriate level of intervention. Conversely, a student with a very low prediction value near 0.0 should receive more aggressive intervention.

The results yielded a very promising 78% accuracy, with the typical caveats of binary classification as described above. The LSTM network used a sequence size of 100 based on the average sequence length. The sequences used actions right up until a pass/fail event. In real-world systems, prediction would need to be based on earlier sequence data to provide timely intervention. This study should continue investigating various windows of sequences between

pass/fail events. Overlapping windows of shorter sequences between events can be tested to see if predictions can be successful while ignoring later data in a sequence. Additionally, more research is needed to investigate if shorter sequences are effective at prediction.

FUTURE RESEARCH

Building on the promising results of this study, several avenues for future research are recommended. First, further investigation into various windows of sequences between pass/fail events could provide deeper insights into the timing and context of at-risk behaviors. Testing overlapping windows of shorter sequences may reveal the effectiveness of predictions made by ignoring later data in a sequence, potentially leading to even earlier detection of at-risk students.

An important direction for future research is the development and testing of interventions based on model predictions. While this study focused on the detection of at-risk students, the next step is to implement and evaluate targeted interventions that can be applied based on the model's predictions. Understanding how different types and intensities of interventions impact student outcomes will be crucial for developing effective support strategies.

Expanding the dataset to include a more diverse range of student interactions and outcomes can enhance the model's generalizability. Including additional features such as demographic information, prior academic performance, and engagement with non-academic resources could further improve the model's ability to predict at-risk students across various educational contexts.

Finally, integrating the predictive model into real-world educational systems and conducting longitudinal studies will provide valuable insights into its practical applications and long-term benefits. By continuously monitoring and refining the model based on real-world data and feedback, researchers can ensure that it remains relevant and effective in supporting at-risk students.

REFERENCES

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539. https://doi.org/10.1109/access.2021.3049446

Al-azazi, Fatima Ahmed, and Mossa Ghurab. "Ann-LSTM: A Deep Learning Model for Early Student Performance Prediction in MOOC." Heliyon, vol. 9, no. 4, 2023, https://doi.org/10.1016/j.heliyon.2023.e15382.

Al Breiki, B., Zaki, N., & Mohamed, E. A. (2019). Using educational data mining techniques to predict student performance. *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*. https://doi.org/10.1109/icecta48151.2019.8959676

Aljaloud, Abdulaziz Salamah, et al. "A Deep Learning Model to Predict Student Learning Outcomes in LMS Using CNN and LSTM." IEEE Access, vol. 10, 2022, pp. 85255–85265, https://doi.org/10.1109/access.2022.3196784.

Bilitski, et. al. " A Machine Learning Approach for Identifying At-Risk Students in Learning Record Stores: A Case Study Using USALearning Experience API (xAPI), IITSEC, 2023.

Buschetto Macarini, L. A., Cechinel, C., Batista Machado, M. F., Faria Culmant Ramos, V., & Munoz, R. (2019). Predicting students success in blended learning—evaluating different interactions inside learning management systems. *Applied Sciences*, *9*(24), 5523. https://doi.org/10.3390/app9245523

Chui, K. T., Fung, D. C., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior*, *107*, 105584. https://doi.org/10.1016/j.chb.2018.06.032

Er, E. (2012). Identifying at-risk students using machine learning techniques: A case study with is 100. *International Journal of Machine Learning and Computing*, 476–480. https://doi.org/10.7763/ijmlc.2012.v2.171

Experience API (xAPI) Specification. (2014) ADL Initiative. https://github.com/adlnet/xAPI-Spec

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using Machine Learning and Explainable AI. *Education and Information Technologies*, *27*(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6

Kondo, N., Okubo, M., & Hatanaka, T. (2017). Early detection of at-risk students using machine learning based on LMS Log Data. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. https://doi.org/10.1109/iiai-aai.2017.51

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2783258.2788620

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2018). Predicting secondary school students' performance utilizing a semi-supervised Learning Approach. Journal of Educational Computing Research, 57(2), 448–470. https://doi.org/10.1177/0735633117752614

Pek, R. Z., Ozyer, S. T., Elhage, T., Ozyer, T., & Alhajj, R. (2023). The role of machine learning in identifying students at-risk and minimizing failure. *IEEE Access*, *11*, 1224–1243. https://doi.org/10.1109/access.2022.3232984

Pilotti, M. A., Nazeeruddin, E., Nazeeruddin, M., Daqqa, I., Abdelsalam, H., & Abdullah, M. (2022). Is initial performance in a course informative? machine learning algorithms as AIDS for the early detection of at-risk students. *Electronics*, *11*(13), 2057. https://doi.org/10.3390/electronics11132057

Soobramoney, Ranjin. *Early Prediction of Students at Risk in a Virtual Learning Environment Using Ensemble Machine Learning Techniques*, https://doi.org/10.51415/10321/4072

Waheed, Hajra, Early prediction of learners at risk in self-paced education: A neural network approach. Exper Systems with Application, 2023.