Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #49083

Beyond Descriptive Statistics: Analysis of Time Series of Student Interactions in Engineering Courses Through the Lens of Modern Mathematical Methods

Dr. Pablo Robles-Granda, University of Illinois at Urbana - Champaign

Pablo Robles-Granda is a Teaching Assistant Professor at the University of Illinois at Urbana-Champaign.

Dr. Hongye Liu, University of Illinois at Urbana - Champaign

Hongye Liu is a Teaching Assistant Professor in the Dept. of Computer Science in UIUC. She is interested in education research based on the Universal Design of Learning framework to help students with disability and broaden participation in computer science.

Celina Anwar, University of Illinois at Urbana - Champaign

Celina Anwar is an undergraduate student majoring in Computer Science. She is currently conducting research under Dr. Pablo Robles-Granda and is interested in the intersection of machine learning, health, and education.

Shivi Narang, University of Illinois at Urbana - Champaign

Shivi Narang is an undergraduate student pursuing Computer Science and Bioengineering at the University of Illinois at Urbana-Champaign. She is currently involved in research with Dr. Pablo Robles Granda, exploring how machine learning can be applied at the crossroads of healthcare and education.

David Dalpiaz, University of Illinois Urbana-Champaign
Prof. Lawrence Angrave, University of Illinois Urbana-Champaign

Dr. Lawrence Angrave is an award-winning computer science Teaching Professor at the University of Illinois Urbana-Champaign. He creates and researches new opportunities for accessible and inclusive equitable education.

Beyond Descriptive Statistics: Analysis of Time Series of Student Interactions in Engineering Courses Through the Lens of Modern Mathematical Methods

Abstract

This study examines the effectiveness of Universal Design for Learning (UDL) tools in engineering courses by analyzing student interaction time series data through statistical and machine learning methods. The primary objectives are to determine (1) whether student engagement with UDL tools is self-informative and (2) to assess whether these interactions can be used to detect engagement changes. Two key UDL components are studied: (a) digital forms, which facilitate non-graded participation and formative feedback, and (b) multimedia tools that provide accessible, self-paced learning opportunities. Student interactions are analyzed using auto-regressive models, including ARIMA, SARIMA, and advanced machine learning methods like GRU and CatBoost. The study also employs Pruned Exact Linear Time (PELT) to detect significant engagement shifts. Findings suggest that student interaction data predicts future engagement, with GRU performing best in minimizing absolute errors and ARIMA excelling in proportional error estimation. Segmentation using PELT enhances predictive accuracy by identifying behavioral shifts. This study shows that classroom-based interactions provide more stable metrics than outside-classroom activities. Ultimately, these methods can help educators improve course accessibility, personalize interventions, and optimize UDL strategies at scale.

Introduction

This study examines the implementation and outcomes of Universal Design for Learning (UDL) activities conducted during the Fall 2023, Spring 2024, and Fall 2024 semesters in three advanced engineering courses at the University of Illinois. This research team has previously developed some of the UDL principles and tools and introduced practices and strategies for content delivery. The approach combined direct student interpersonal collaboration, behavior, and perspective, leveraging in-class UDL interaction measures and outside-class UDL use. The primary goal of this article is to provide a case study for the ASEE community and engineering educators by analyzing two key UDL strategies: 1) encouraging student participation with in-class UDL tools and 2) fostering knowledge internalization via out-of-classroom UDL tools. To evaluate the effectiveness of these strategies, we developed our student interaction metrics based on traffic and interaction data we collected from these tools. Our findings indicate that the distribution of concise, UDL-based evaluation of course activities positively impacts students' performance, with engagement levels varying by course section and semester timing. This was evidenced by the growth/decrease in interaction activity and the expected performance. The study also applies statistical and machine learning models to analyze interaction patterns and predict future interactions. We also model changes in student interactions and report the best-performing models for this task. Outside-class and inside-classroom interactions were analyzed and validated to assess differences in behaviors. This study highlights the potential of UDL strategies to improve student interaction in advanced engineering courses, providing insights for educators seeking data-driven instructional analysis.

Background

Universal Design for Learning (UDL) is a comprehensive educational framework that promotes inclusivity by adopting strategies to support diverse student learning, expression, and engagement methods, ultimately enhancing academic outcomes. Three fundamental principles guide this approach: 1) presenting content through various methods to accommodate different learning preferences, 2) enabling students to demonstrate their understanding through multiple forms of expression, and 3) fostering engagement and motivation through diverse means. These practices ensure equitable access to learning opportunities for all students, including those with disabilities (SWD) [1]. UDL strategies include onboarding forms, frequent low-stakes assessments, and flexible assignment deadlines. While much of the existing UDL work focuses on evaluating its effectiveness and developing innovative tools, this paper addresses the challenge of applying these tools in advanced engineering courses and tracking changes from a student perspective.

Measuring Educational Effectiveness of Accessibility in Advanced Engineering Content: The accessibility of upper-level university engineering courses is a pressing concern, with Universal Design for Learning (UDL) emerging as a pivotal framework for fostering inclusivity. UDL principles, which advocate for multiple means of representation, action, and engagement, aim to minimize barriers for diverse learners, including students with disabilities (SWDs). Key challenges include resistance to adopting inclusive practices and technological constraints within Learning Management Systems (LMS). These issues disproportionately affect SWDs and students from underrepresented groups, often hindering their full participation [2–4]

Effective implementation of UDL-guided course designs promotes equitable learning outcomes. For instance, hybrid and asynchronous courses employing multimodal teaching strategies—such as visual, auditory, and interactive content—demonstrate improved accessibility and deeper student engagement [2, 5, 6]. However, significant gaps remain in the consistent application of UDL principles. Studies highlight that faculty often lack awareness or resources to adapt their teaching practices, leading to fragmented efforts in addressing accessibility challenges [2, 4].

Engineering education uniquely benefits from integrating UDL and inclusive design principles. Inclusive design projects, like creating assistive tools for individuals with disabilities, foster empathy, innovation, and real-world problem-solving skills among students [2, 7]. Such projects resonate particularly with underrepresented groups, including SWDs, who are motivated by the societal impact of their work. Capstone courses, for example, effectively incorporate UDL to encourage students to consider diverse user needs in their designs [2].

Despite these advancements, SWDs report significant barriers, including difficulties navigating multiple LMS platforms, inconsistent use of accessible tools, and limited instructor awareness. Surveys reveal that centralized platforms, captioned videos, flexible deadlines, and unified course calendars significantly enhance accessibility and engagement [3, 7]. However, many SWDs refrain from disclosing their disabilities due to many reasons, including stigma or distrust in receiving timely accommodations [6, 7]. Thus, accessibility should be prioritized in course tools, without making assumptions about students' disability status. In this paper, we focus on the use of two types of tools: 1) mobile-device-friendly student tools that can facilitate student interactions due to the ease of access and the fact that smartphone UX tends to be significantly better than a computer, and 2) multimedia recordings and transcripts that student can use at their own pace.

UDL Guidance for Faculty. Faculty expertise on UDL and accessibility is crucial to overcoming resistance and fostering learning. Programs that train instructors to leverage adaptive technologies and implement UDL principles effectively enhance course accessibility and student satisfaction [3, 4]. Using features like adaptive assessments, multimodal content, and real-time feedback significantly improves the learning experience for diverse students [3, 5]. The potential of UDL extends beyond accessibility for SWDs, benefiting all students by creating flexible, engaging learning environments because UDL-based pedagogies seem to improve retention rates, reduce dropout rates, and attract a more diverse student body [2, 5]. Thus, systemic and comprehensive faculty training is needed for effective teaching [2, 5, 7]. Integrating UDL principles into upper-level engineering courses could not only remove barriers for SWDs but also enrich the educational experience for all learners. Engineering programs can better prepare students by aligning educational practices with student needs [2]. In this paper, we report lessons learned from applying UDL techniques inside and outside the classroom as part of the usual strategies for two courses at the University of Illinois.

Methods

This section describes two key elements of our work. First, we describe the UDL tools used to engage student participation inside and outside class. This corresponds to anonymized time-varying data. Second, we describe the process we follow to extract variables that can serve to identify the level of engagement the students have with the course materials in the upper-level course under study. The former includes details of the three-fold view of UDL: the modalities to accommodate learning preferences, multiple forms of expression, and varying types of engagement of students. The latter includes a description of the model's rationale, structure, and results. We also provide detailed implementation information for our methods, including assumptions and scalability considerations.

UDL Tools Applied

We applied various UDL strategies in the courses under study, including daily lesson goals, flexible workspaces (including varying sizes of student interactions) both in person and through forums, and multimedia material (including videos and digital whiteboards). In this report, we will focus on the tools and activities listed below, which were later used to understand the students' responses and perspectives. The objective of using the UDL strategies below is to track student activity without assigning grades. Non-graded UDL activities and assignments could foster an inclusive, low-pressure learning environment emphasizing growth over performance from various perspectives. First, it could encourage a growth mindset, reduce stress, and promote intrinsic motivation by allowing students to focus on mastering skills without fear of failure. Second, this approach supports self-regulation and reflection, enabling students to take ownership of their learning. By focusing on formative assessment and providing meaningful feedback, teachers can address diverse learning needs and create opportunities for individualized support. Third, it encourages collaboration, peer learning, and creativity while reducing anxiety tied to grades. Tracking activities helps teachers adapt instruction and identify learning gaps, aligning with mastery-based learning principles. Finally, students are empowered to take risks, engage deeply, and develop a love for learning. This approach builds a supportive classroom culture where all students can succeed. We focus on two components.

Component A. Digital Forms. Online tools like Google Forms and Polls Everywhere align with UDL principles by offering flexible, inclusive, and interactive learning opportunities. These tools provide multiple means of representation by presenting information in diverse formats such as text, visuals, or videos, making content accessible to all learners. They enhance engagement by encouraging active participation through polls, quizzes, and surveys, with options for anonymity to reduce anxiety. Students can express their understanding in various ways, accommodating diverse communication styles and preferences. More precisely, we use digital forms for student activities that are not graded for completion but for participation. We follow a structured format for the interaction for most of the forms, but allow a few to be answered via pencil and paper, which is later uploaded as an image. All of the forms allow for open-ended questions that students can answer with no constraint on the number of characters, but then are summarized or addressed during class. The structure of the form is concrete and is as follows:

Element	**Type**
Lecture Semester and Number	Label
Form's Topic	Label
Question(s)	Field (text/image/multiple choice - several combined)

Table 1: Elements of the Course Interaction Tools

Component B. Multimedia Material (ClassTranscribe). Online multimedia tools like ClassTranscribe provide inclusive, accessible, and flexible learning opportunities. By automatically generating transcripts from lecture videos, ClassTranscribe ensures multiple means of representation, making content accessible to students with hearing impairments, language barriers, or different learning preferences. Transcripts also allow students to interact with the material in a textual format, enabling them to search, highlight, and review key concepts at their own pace. This gives students control over how and when they access content, such as revisiting lectures for better understanding or studying asynchronously. It also supports multiple means of action and expression by allowing students to integrate video and text-based content into their learning, giving students the format that best suits their needs. In the courses under study, students are allowed not only to use and engage with the recordings of the onsite lectures via ClassTranscribe but also have the opportunity to contribute to other students by submitting corrections to errors in the transcription.

Research Questions. To evaluate the methods above, we focus on two elements: namely, how self-informative the interaction of students with UDL tools is, and second, how to use these interactions to automatically identify changes in student usage of UDL tools that can be used later to determine appropriate measures to help students. Thus, in the remainder of this paper, the accessibility of engineering courses is analyzed via two research questions:

RQ1: Is student interaction with UDL tools self-informative?

RQ2: Can student interaction be used to assess changes in engagement?

Notice that "self-informative" in this context means that student interactions with UDL tools could predict future performance, engagement, or learning outcomes without needing additional external factors. Thus, to answer RQ1, we select models that show how past interaction data predict future engagement levels and how interactions' frequency, type, or quality correlate with self-reported understanding or learning outcomes. To answer the second question (RQ2), we aim to identify changes using statistical and machine learning auto-regressive methods by studying how changes in engagement are reflected in time-series interaction data.

Model Specification and Evaluation

To answer both research questions, we will consider auto-regressive models. Auto-regressive models are highly appropriate for analyzing Universal Design for Learning (UDL) data because they account for temporal dependencies, repeated measures, and hierarchical structures inherent in student engagement data. Engagement patterns, such as tool usage or quiz completion, are influenced by prior behaviors, making AR models ideal for capturing these dynamics. These models can track how past interactions with UDL tools (e.g., time spent on videos) predict future engagement, allowing educators to identify trends, cycles, or abrupt changes. Unlike simpler models, AR methods handle lagged effects, showing how long past behaviors impact future outcomes, which is crucial for understanding sustained engagement. AR models can also be extended to multi-variable contexts (e.g., Vector Auto-Regression) to explore how different UDL activities, like video watching and quiz attempts, interact over time. They support forecasting, enabling educators to predict engagement changes and intervene proactively. By identifying systematic patterns, AR models distinguish between meaningful trends and random noise, providing actionable insights. Extensions like ARIMA or machine learning-based LSTMs enhance the ability to handle non-linear, seasonal, or multi-scale engagement behaviors. This makes auto-regressive modeling a powerful tool for automating the assessment of student engagement and improving UDL strategies over time. The steps for our analysis are:

Step-0. Data preprocessing - Data preprocessing was performed in two steps. First, the variable used for measuring engagement was the response length for the student interactions. This is based on the analysis of the responses where shorter texts indicated less interest in the topic (e.g., "I don't know" answers). This does not preclude the potential use of other summary statistics to describe the data, including semantic analysis. Second, the data imputation was implemented using the within-subject median for missing values (e.g., NaN). We allow zero-length responses precisely because we are interested in measuring student-initiated contributions. These steps were performed on anonymized data, so no access to additional variables was possible. After these steps, we formatted the data as a time series for further processing without normalization.

Step-1. AR Model Selection (To address RQ1) - As detailed before, we applied statistical and machine learning autoregressive models for the benefits listed below, which can make them suitable for educational data. Specifically, we studied:

1. Auto-Regressive Integrated Moving Average (ARIMA). ARIMA is excellent for modeling univariate time-series data with temporal dependencies, especially when the data is stationary or can be transformed into stationarity.
2. Seasonal Auto-Regressive Integrated Moving Average (SARIMA). SARIMA extends ARIMA by incorporating seasonality, making it ideal for time-series data with repeating

patterns (e.g., weekly or monthly cycles). It can capture both short-term trends and long-term periodicity.

3. Exponential Smoothing with Holt-Winters. Holt-Winters is suitable for data with seasonality and trends. It uses exponential smoothing to weigh recent observations more heavily.
4. Extreme Gradient Boosting (XGBoost). XGBoost is a robust machine-learning algorithm that can handle non-linear relationships in time-series data by leveraging gradient boosting on decision trees. It is highly efficient for large datasets and supports missing data.
5. CatBoost. CatBoost is ideal for time-series data with categorical features, as it efficiently handles categorical variables without preprocessing.
6. Long Short-Term Memory (LSTM) LSTMs excel in modeling long-term dependencies in sequential data, capturing complex patterns that traditional models might miss. Their memory cell architecture is particularly effective for handling time-series data with non-linear relations.
7. Gated Recurrent Unit (GRU). GRUs are a more straightforward and computationally efficient alternative to LSTMs. They offer comparable performance in modeling sequential data with short- and medium-term dependencies and for faster training and inference.

Step-2. Identification of Changes in Engagement (To answer RQ2) We used Pruned Exact Linear Time (PELT) to study variations in the student activities. PELT is a fast and accurate change-point detection algorithm that identifies significant shifts in time series data, such as trends, variance, or mean changes. It is computationally efficient and scalable, making it well-suited for detecting engagement changes in real time for scenarios where educators intend to apply interventions in their courses on the fly.

Step-3. Example of benefits: Re-evaluation of AR Models on the Identified Change Points. (Application of Step-2 to improve Step-1). We provide an example of how to use *Step-2* to design further analysis of course designs. Note that this is not the only application, and other decisions can be made based on the Step-1 analysis, such as course interventions. Change-point detection is valuable for partitioning time-series data in UDL contexts by identifying when significant shifts in student engagement occur, such as sudden drops in tool usage or spikes in activity. This segmentation improves the accuracy of auto-regressive models by allowing them to focus on stable engagement patterns, reducing noise in predictions. It also enables educators to design timely, targeted interventions based on the timing and nature of detected changes, improving the effectiveness of UDL tools. By analyzing separate periods, change-point detection enhances feature engineering, capturing localized trends and seasonal components more effectively. Additionally, it helps uncover underlying causes of engagement shifts, such as curriculum updates or external events, allowing UDL strategies to adapt to students' needs. These insights also support efficient resource allocation by ensuring interventions focus on periods of greatest impact, ultimately optimizing student engagement and learning outcomes.

Model Formula – Observation: The solution of *Step-1* can be used to identify a function $x_t = f(x_{t-1}, \ldots, x_{t-n})$ that links the multivariate variables of student interactions, x_t , based on historical information $x_{t-1}, \ldots, x_{t-n}$ This relation is not necessarily linear, and its complexity can be addressed via auto-regressive models that can capture these nonlinear relations.

Assumptions and Diagnostics:

Our analysis assumes that the length or duration of student interactions with course elements (or UDL components) can provide information that may be correlated with students' interests in the course content. We also assume that this interaction is highly variable (due to changes in the course pace and the student's interests) and nonlinear. For this reason, we apply both parametric and non-parametric models to run our predictions.

Scalability:

The final selected model, in our case, would be the only scaling constraint for the recommendation in this paper. ARIMA and SARIMA are computationally efficient and scalable for small to moderately sized time-series datasets. Still, they can struggle with extensive or high-dimensional data due to their reliance on manual parameter tuning. Holt-Winters is lightweight and highly scalable for simple time-series data, notably when handling trends and seasonality, but its performance diminishes for complex or non-linear patterns. XGBoost and CatBoost are highly scalable for large datasets, leveraging parallel processing and efficient memory usage, making them ideal for high-dimensional and categorical data. LSTM and GRU, while powerful for capturing complex and non-linear temporal dependencies, are computationally intensive and require significant resources for training on large datasets. However, GRU is slightly faster due to its simpler architecture. PELT, as a change-point detection algorithm, is highly scalable and efficient, operating in linear time, making it suitable for large time-series datasets where detecting shifts in trends or patterns is critical. These models offer varying levels of scalability, with simpler statistical models excelling in smaller datasets.

Results

Part I - Data Description

The datasets we study in this paper include students' interactions for two upper-class engineering courses with a strong programming and mathematical focus. The statistics of these courses are listed in Table 2.

Course	Group	Number of Subjects
Course I	A	96
Course I	B	47
Course II	A	94
Course II	B	50
Total		287

Table 2: Number of subjects per group and course

The main difference between courses I and II behaviors is the interactions of time stamps 25-27 for Course I, where the level of activity was greater. While Course II was larger, the number of subjects listed in the table for each course is approximately the same for our analysis because Course II also had students who enrolled in the online version of the course.

Part II - Prediction Models

Autoregressive Analysis.

We applied statistical and machine learning tools to assess the self-informative patterns of student activities. The models in these results are listed in the previous section, and the acronyms are listed in the appendix.

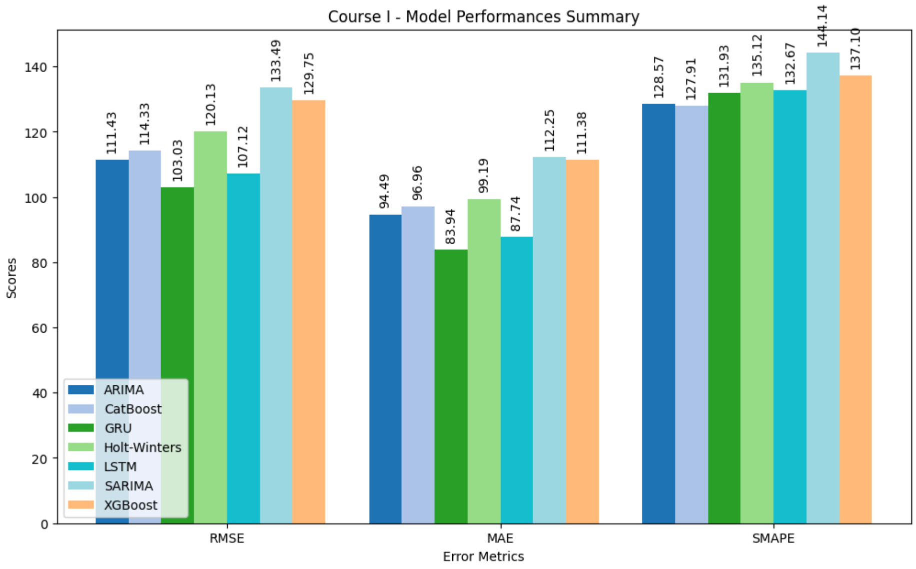

Figure 1: Course I – Model Performances Summary (averaged across all subjects)

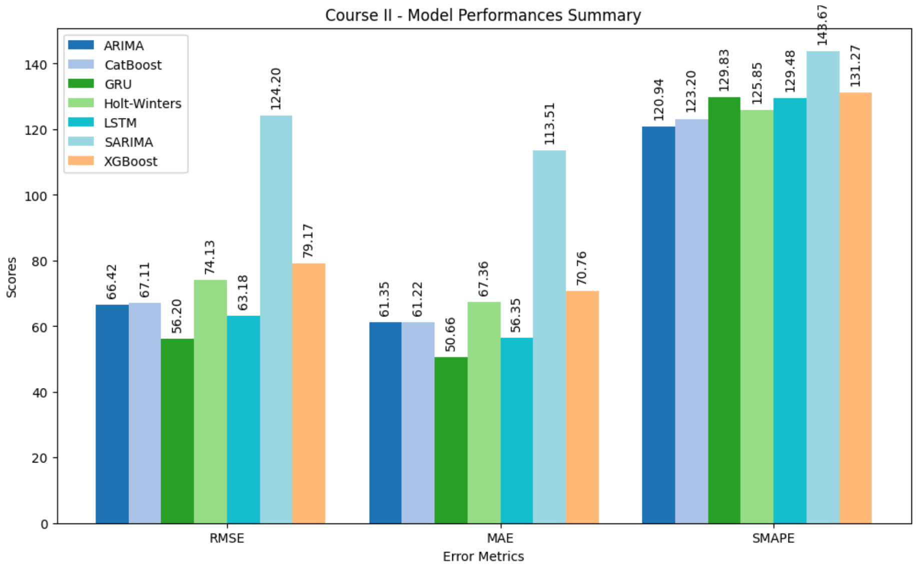

Figure 2: Course II – Model Performances Summary (averaged across all subjects)

Figures 1 and 2 summarize the performance of the models for both Course I and Course II. The figures compare the performance of the various models we introduced in the Methods Section based on three evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (SMAPE). These metrics measure the accuracy of each model, where lower values indicate better performance. Figure 1 shows slightly higher errors for the first course than the second, suggesting the second course might have less variability or nonlinearities. Errors (RMSE, MAE, SMAPE) are generally lower in Figure 2, indicating that the second dataset is likely less complex or shows less variability, where models show better performance overall, with smaller error differences.

Best Performing Model (Overall Lowest Errors): As shown in Figure 1, GRU and CatBoost are the best-performing models – GRU achieves the lowest RMSE (103.03) and MAE (83.94), while CatBoost achieves the best SMAPE (127.91). In Figure 2, GRU has the lowest RMSE (56.20) and MAE (50.66), indicating consistent performance across datasets. CatBoost performs slightly worse than GRU but still shows good SMAPE.

As seen in these figures, GRU is the most robust model across both datasets, achieving the lowest RMSE and MAE consistently, making it the best option for accurate predictions; CatBoost is the best model for handling proportional errors (SMAPE), especially in the second dataset. Statistical models like ARIMA perform well compared to other statistical methods (SARIMA and

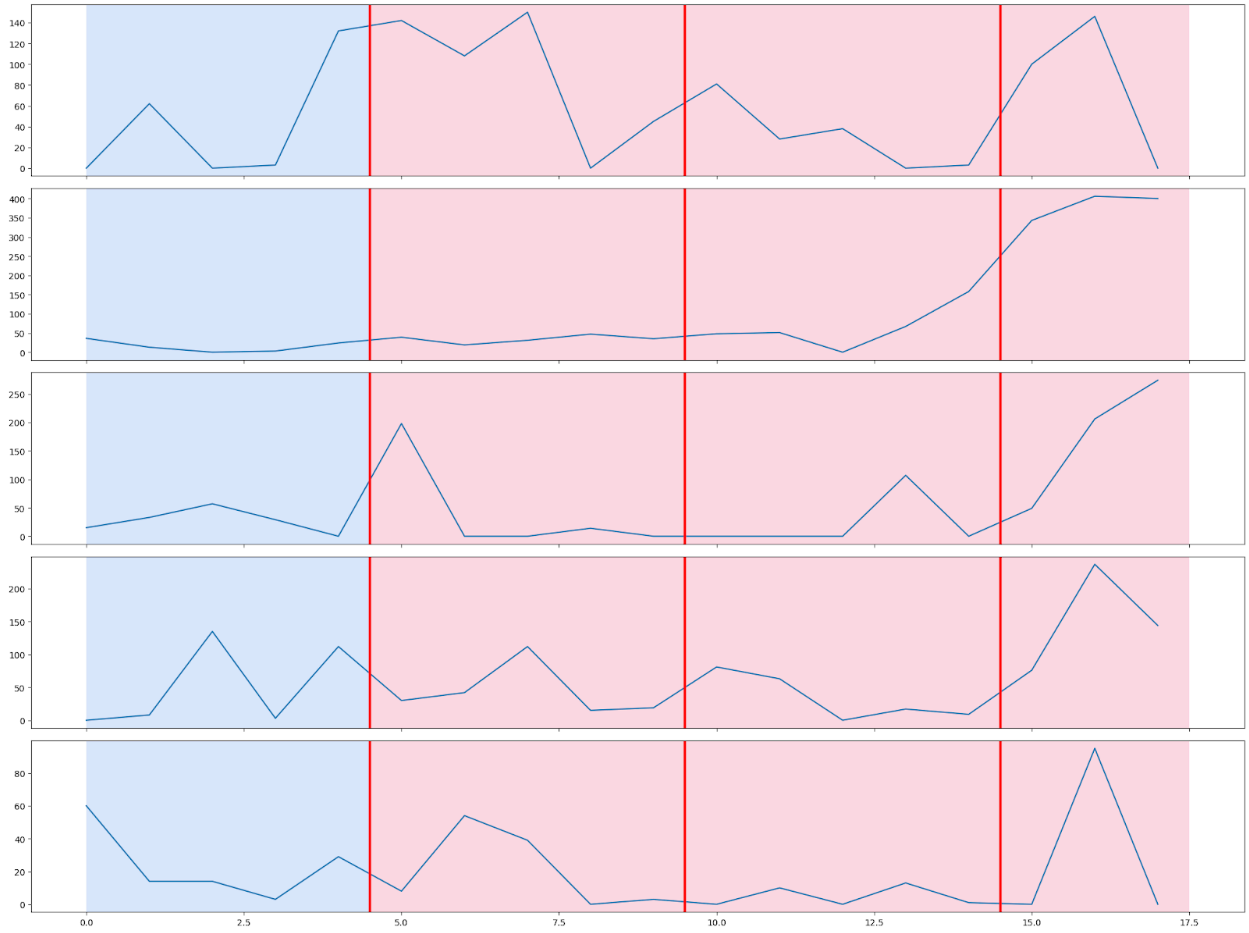

Figure 3: Course I Partitioning – Sample of 5 interactions

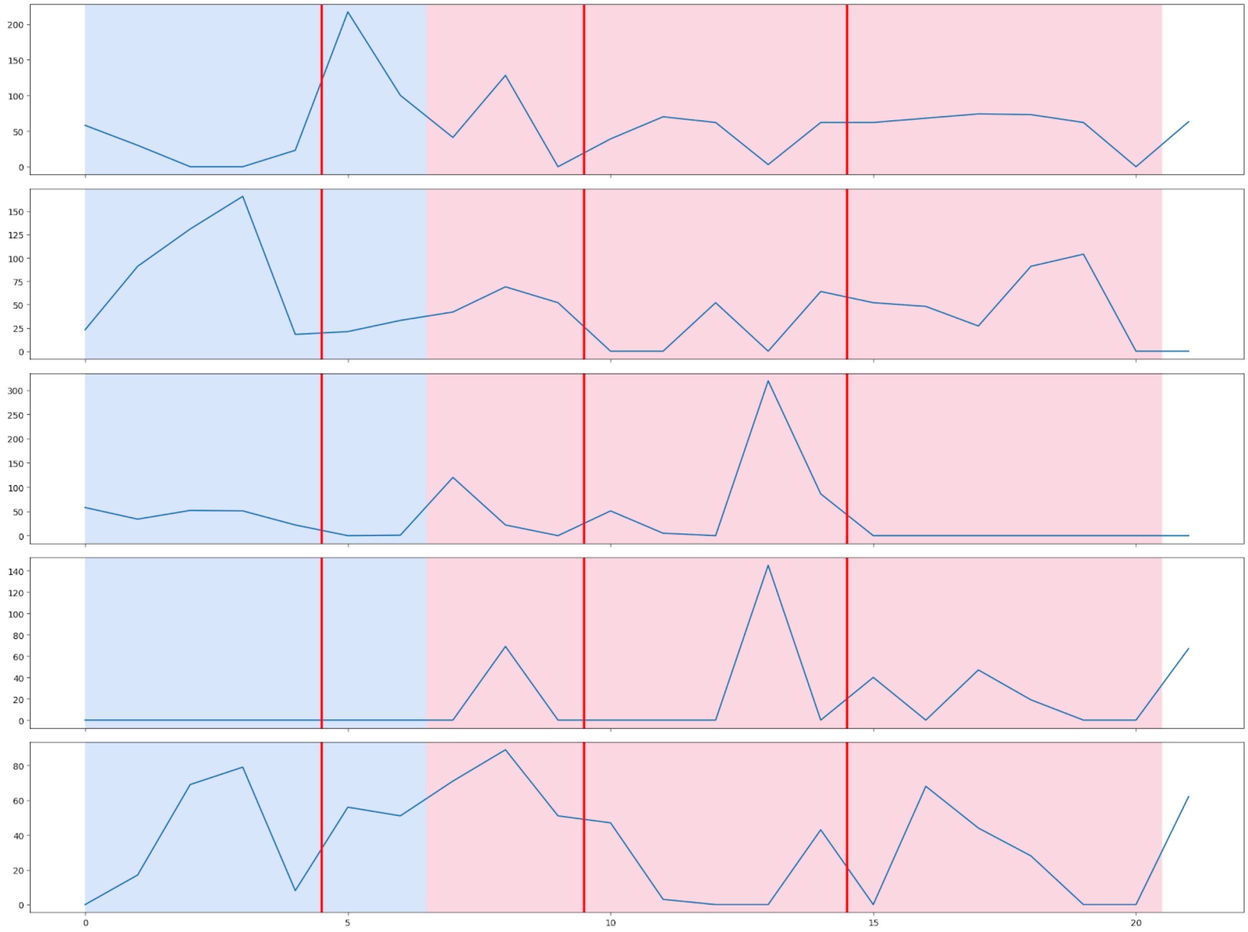

Figure 4: Course II Partitioning – Sample of 5 interactions

Holt-Winters) but fall behind machine learning (GRU and CatBoost) in both datasets. Machine learning models (CatBoost, XGBoost) perform well on more straightforward longitudinal datasets but are outperformed by deep learning models (GRU, LSTM) for complex temporal patterns, as our analysis corroborates. SARIMA is the lowest-performing model across both datasets.

Part III - Behavior Shifts

Understanding Variation of Engagement.

We run a segmentation analysis for the time series of engagement data to understand sudden changes in student engagement. Figures 3 and 4 show the results of this analysis for Course I and II, respectively, for a selection of 5 subjects. The remaining subjects have the same estimated partition of their time series, as this is a between-subject segmentation of the time series. The figures show the times before the first evaluation of the semester; the time series are color-coded before (blue) and after the assessment (red). As shown there, there are three main points identified by the PELT method with *l1* loss.

Predicting student engagement with the segmented series:

In this section, we compare the performance of two models (ARIMA and GRU) on the time series data by partitioning the training and test sets on the segments identified by the PELT method.

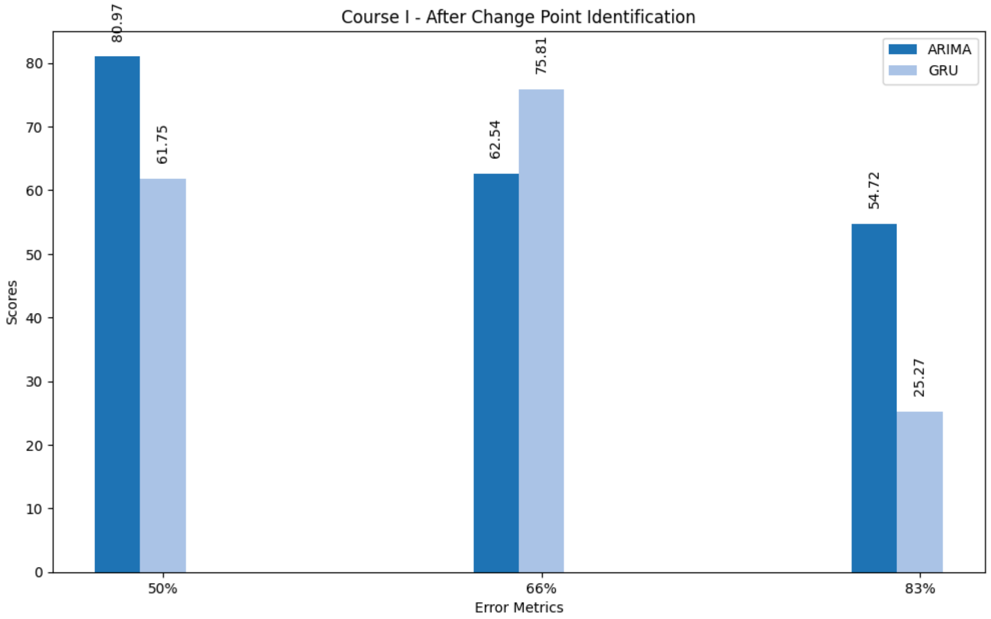

Figure 5: Course I – After Change Point Identification – Model Evaluation Summary (averaged across subjects). Errors decrease with increasing data sizes.

Thus, we used datasets of varying sizes and proportions, evaluated using three metrics: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and SMAPE (Symmetric Mean Absolute Percentage Error). For illustrative purposes, Figure 5 shows the results for the RMSE error, while Table 3 (Appendix) shows the details for the other error metrics. In Figure 5, each bar corresponds to a different dataset size (50%, 66%, and 83%), where the percentages represent how much of the total data each dataset constitutes.

As shown in Figure 5 (and Table 5 in the Appendix), GRU is the best model for minimizing absolute errors (RMSE, MAE), particularly on small or relatively larger datasets, and shows significant improvement as data size grows. On the other hand, ARIMA consistently handles proportional errors (SMAPE) better and performs well on medium-sized datasets for RMSE and MAE. Thus, our general recommendation for instructors is to use GRU when reducing absolute prediction errors is a priority, especially with relatively larger datasets, and use ARIMA for applications requiring proportional error minimization or when data size is moderate. We discuss some specific and detailed technical observations in the Appendix.

Trends Across Datasets Per Error and Data Size: GRU generally outperforms ARIMA in minimizing absolute errors (RMSE and MAE), especially in non-trivial or larger datasets. ARIMA performs better in moderate dataset volumes. For proportional errors, ARIMA performs better. More technical details on these analyses are presented in the appendix.

Sudden Changes in Student Interaction: We analyzed the time-series segmentation presented in this section to identify sudden changes in student interactions. In the Appendix, we present the details of this analysis, showing the segmentation points where the prediction error is minimized. Table 4 shows the results of identifying sudden changes in student interactions, with lower error values observed at segmentation points 5, 10, and 14 for data sizes 10, 15, and 22.

Outside of the Classroom Interaction: We analyzed student interactions with course materials outside the classroom by counting daily interactions. Fewer students engaged with the tool, with only 44 participating and 20 doing so for three or fewer days. GRU slightly outperforms LSTM across all evaluation metrics when averaged across subjects. Performance worsens as the training ratio increases, suggesting that both models struggle with larger datasets. However, GRU performs better in SMAPE, indicating its advantage when considering proportional errors and normalized data. Further technical details of this analysis are presented in the Appendix.

Conclusion

In this work, we analyzed two specific questions that educators face when designing and monitoring the accessibility of advanced engineering courses, namely, how self-informative the information in UDL tools is and whether student interaction/UDL data can be used to assess changes in engagement. Based on our analysis, the answer to both questions seems affirmative. First, student interaction with UDL tools is self-informative, as evidenced by the ability of models like GRU to predict future interaction patterns. Second, student interaction data can be used to assess changes in engagement, as segmentation methods like PELT effectively identify shifts in behavior, enabling tailored predictive modeling.

To answer the research questions, our analysis applies statistical and machine learning tools to assess student activity patterns and predict UDL strategy engagement using error metrics such as RMSE, MAE, and SMAPE. GRU is the most robust model across datasets, consistently achieving the lowest RMSE and MAE, making it highly accurate for minimizing absolute errors. CatBoost, however, performs better for SMAPE, indicating its strength in handling proportional errors, particularly in simpler datasets. ARIMA and other statistical models like SARIMA perform reasonably well but lag behind GRU and CatBoost.

Segmentation analysis using the PELT method highlights shifts in student engagement, revealing partitions that help optimize model performance by training and testing within these segments. GRU outperforms ARIMA for absolute error minimization on smaller (size 10) and larger datasets (size 18), while ARIMA performs better on medium-sized datasets (size 15), particularly for SMAPE. These results suggest GRU excels with more data, while ARIMA is more stable across varying dataset complexities.

GRU and LSTM exhibit lower variance across all metrics, making them more reliable for consistent predictions than statistical models like ARIMA and SARIMA. GRU also has the tightest distribution for RMSE and MAE, confirming its consistency. In conclusion, GRU is recommended for absolute error minimization, especially with larger datasets, while ARIMA is better suited for proportional error reduction (SMAPE) or moderate data sizes. CatBoost can serve as an alternative for proportional errors in some instances.

The results for inside-the-classroom metrics are more consistent than outside-classroom due to several factors, including the size of the data (fewer elements in the outside-classroom case), the length of the time series (fewer timestamps and different lengths for outside-classroom), and other behavioral elements (different motivations or opportunities may lead students to interact inside and outside the classroom differently). These findings highlight the importance of model selection and error prioritization in UDL applications.

Discussion

Our experience shows that digital forms and polls support low-stakes feedback, allowing students to engage without fear of failure and fostering a safe and growth-oriented learning environment. Features like accessibility options and real-time feedback ensure inclusivity for students with disabilities or specific needs. Additionally, they promote self-reflection and peer learning by encouraging students to review their responses and compare them with others. By incorporating these tools, educators can create dynamic, student-centered learning experiences that cater to various needs and preferences. Likewise, multimedia tools, such as ClassTranscribe, help with self-regulation by assisting students to track their progress and manage their learning independently. It also provides equitable access for students who may not be able to attend live lectures or need additional time to process information. By incorporating such tools, educators can create a more inclusive and adaptable learning environment that meets the diverse needs of all students and aligns with UDL principles. Regarding scalability, the models and strategies presented in this paper can be applied to large courses (of about hundreds of students). Other methods must be considered to scale the analysis to larger settings (such as massive open online courses) where centering predictors, standardizing continuous variables, and normalization could be used to ensure scalability.

In addition to the various models tested in our study, including ARIMA, SARIMA, and machine learning models like GRU and CatBoost, we tested the Holt-Winters method, which can provide valuable insights. Holt-Winters represents time series with explicit seasonal patterns through its three-component structure (level, trend, and seasonality). However, for the UDL interaction data in our advanced engineering courses, Holt-Winters may be less appropriate than the selected models, which likely stems from the complex, non-linear patterns in student engagement that don't align perfectly with Holt-Winters' assumptions of consistent seasonal cycles. In particular, Holt-Winters struggles to adapt to abrupt behavioral shifts or irregular usage patterns, which are common in educational settings driven by deadlines or exams. Since it cannot model non-seasonal variance well, it tends to underfit in contexts where interaction frequency changes unpredictably. As shown in our results, machine learning models like GRU outperformed traditional statistical approaches for minimizing absolute errors, suggesting that capturing the nuanced relationships in educational interaction data requires more flexible models.

Our general recommendation for instructors is to use GRU when reducing absolute prediction errors is a priority, especially with larger UDL datasets, and use ARIMA for applications requiring proportional error minimization or when data size is moderate. However, time series with fewer than ten elements or five timestamps can lead to issues, and traditional statistical methods may be preferred in those scenarios.

Acknowledgments

We thank the National Center for Supercomputing Applications - NCSA, particularly Rob Kooper, Lead Research Software Engineer at NCSA, for technical assistance and support of ClassTranscribe. C. Anwar and S. Narang contributed equally to this work.

References

[1] E. Scanlon, J. Schreffler, W. James, E. Vasquez, and J. J. Chini, "Postsecondary physics curricula and universal design for learning: Planning for diverse learners," *Physical Review Physics Education Research*, vol. 14, no. 2, p. 020101, 2018.

[2] B. Blaser, K. M. Steele, and S. E. Burgstahler, "Including universal design in engineering courses to attract diverse students," in *2015 ASEE Annual Conference & Exposition*, 2015, pp. 26–935.

[3] H. Liu, D. Moparthi, L. Angrave, J. Amos, D. Dalpiaz, C. Vogiatzis, S. Varadhan, Y. Huang, and R. Reck, "Understanding the needs of students with and without disabilities for inclusive UDL-based design of engineering courses through learning management systems," in *2022 ASEE Annual Conference & Exposition*, 2022.

[4] S. Varadhan, X. Ding, D. L. Zhao, A. Agarwal, D. Dalpiaz, C. Vogiatzis, Y. Huang, L. Angrave, and H. Liu, "Opportunities and barriers to UDL-based course designs for inclusive learning in undergraduate engineering and other STEM courses," in *2023 ASEE Annual Conference & Exposition*, 2023.

[5] J. Livingston, S. Summers, and J. Szabo, "Incorporating universal design for learning principles in online and hybrid technical communication courses," *Journal of Online Engineering Education*, vol. 10, no. 2, 2019.

[6] Z. Kang, M. R. Dragoo, L. Yeagle, R. L. Shehab, H. Yuan, L. Ding, and S. G. West, "Adaptive learning pedagogy of universal design for learning (UDL) for multimodal training," *Journal of Aviation/Aerospace Education & Research*, vol. 27, no. 1, pp. 23–48, 2018.

[7] I. C. Landa-Avila and C. Aceves-Gonzalez, "Inclusive human-centered design: experiences and challenges to teaching design engineering students," in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*. Springer, 2019, pp. 1558–1570.

Appendices

A Technical Details and Additional Analysis

Predicting student engagement with the segmented series:

In this section, we provide additional details of the performance comparison between ARIMA and GRU models on the time series data, but partition the training and test sets on the segments identified by the PELT method. Thus, we report results on varying sizes and proportions of the training sets, evaluated using all three metrics: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and SMAPE (Symmetric Mean Absolute Percentage Error). In Table 3, each sub-table corresponds to a different dataset size (50%, 66%, 83%), with percentages representing how much of the total data each dataset constitutes. As shown in Table 3, GRU is the best for minimizing absolute errors (RMSE, MAE), particularly on small or large datasets, and shows significant improvement as data size grows. On the other hand, ARIMA consistently handles proportional errors (SMAPE) better and performs well on medium-sized datasets for RMSE and

RMSE	MAE	SMAPE	
data size and percentage: 10; 50%==5			
ARIMA	80.97	74.39	125.70
GRU	61.75	55.32	169.24
data size and percentage: 15; 66%==10			
ARIMA	62.54	51.90	99.30
GRU	75.81	71.22	154.90
data size and percentage: 15; 66%==10			
ARIMA	54.72	54.72	90.95
GRU	25.27	25.27	142.06

Table 3: Course I - After Change Point Identification - Model Evaluation Summary (averaged across all subjects)

MAE. Thus, our general recommendation for instructors is to use GRU when reducing absolute prediction errors is a priority, especially with larger datasets, and use ARIMA for applications requiring proportional error minimization or when data size is moderate. We note some consistent behavior across training set sizes:

Case 1: Data Size = 10 (0.5 training)

- RMSE and MAE: GRU (RMSE = 61.75, MAE = 55.32) significantly outperforms ARIMA (RMSE = 80.97, MAE = 74.39), demonstrating lower overall and average prediction errors.
- SMAPE: ARIMA achieves better SMAPE (125.70) compared to GRU (169.24), indicating better performance in handling percentage-based errors.
- Observation: On this small dataset, GRU performs better for absolute errors (RMSE, MAE), but ARIMA handles proportional errors more effectively.

Case 2: Data Size = 15 (0.66 training)

- RMSE and MAE: ARIMA (RMSE = 62.54, MAE = 51.89) now outperforms GRU (RMSE 75.81, MAE = 71.22), showing better overall & average accuracy for medium sized data.
- SMAPE: ARIMA (SMAPE = 99.30) remains better than GRU (SMAPE = 154.90), maintaining its advantage in proportional error handling.
- Observation: ARIMA begins to outperform GRU across all metrics with a slightly larger dataset, indicating its robustness on medium-sized data.

Case 3: Data Size = 18 (0.834 training)

- RMSE and MAE: GRU (RMSE = 25.27, MAE = 25.27) substantially outperforms ARIMA (RMSE = 54.72, MAE = 54.72), with much lower prediction errors.
- SMAPE: ARIMA achieves better SMAPE (90.95) than GRU (142.06), showing continued strength in proportional error reduction.
- Observation: On this larger dataset, GRU excels in absolute error metrics, while ARIMA remains more effective for SMAPE.

Trends Across Datasets Per Error:

RMSE vs MAE: GRU performs better than ARIMA on the smallest (10) and largest (18) datasets, suggesting it benefits from smaller or larger data sizes. ARIMA performs better for the medium-sized dataset (15), indicating it may excel with moderate data. SMAPE: ARIMA consistently outperforms GRU in proportional error handling across all dataset sizes. This makes ARIMA more reliable for applications where percentage-based errors are critical.

Impact of Data Size:

GRU significantly improves absolute error metrics (RMSE and MAE) as data size increases, suggesting it leverages more data effectively. ARIMA remains relatively stable but performs best on medium-sized data for absolute errors and handles SMAPE well.

Sudden Changes of Behavior and Error Values:

Table 4 shows an additional analysis of the applicability of time series segmentation to identify sudden changes in student interactions. The error values are lower for segmentation points 5, 10, and 14 (data sizes 10, 15, 22).

RMSE	MAE	SMAPE
datasize=10; 0.5==5 training:		
RMSE	MAE	SMAPE
61.87	51.15	124.18
datasize=15; 0.67==10 training:		
RMSE	MAE	SMAPE
98.01	89.81	136.91
datasize=22; 0.68==14 training:		
RMSE	MAE	SMAPE
96.71	73.37	170.48

Table 4: Course II performance for GRU

Evaluation Outside of the Classroom:

We also analyze the interactions with the materials outside of the classroom. For these, we counted the number of interactions each student had per day. The number of students interacting was smaller, with 44 interacting with the tool, and 20 did it in 3 or fewer days. Table 5 summarizes the performance of GRU and LSTM models across all subjects using four metrics: RMSE, MAE, and SMAPE. GRU slightly outperforms LSTM in all metrics.

Model Evaluation Summary(averaged across all subjects):

model	RMSE	MAE	SMAPE
GRU	363.23	276.40	120.45
LSTM	366.74	284.57	128.94

Table 5: Outside Classroom - Models' Prediction Performances

Table 6 breaks down model performance (GRU and LSTM) across three training ratios: 0.4, 0.5, and 0.6, using the same metrics as Table 5. The models generally exhibit worsening performance

RMSE by Training Ratio and Model:				
Training %	Model	RMSE	MAE	SMAPE
0.4	GRU	330.82	230.45	133.04
	LSTM	329.90	223.30	138.10
0.5	GRU	371.39	283.32	105.85
	LSTM	385.73	321.84	125.09
0.6	GRU	387.47	315.42	122.45
	LSTM	384.59	308.58	123.64

Table 6: Outside Classroom - Models' Prediction Performances Per Error Rate

(higher RMSE and MAE) as the training ratio increases, indicating the models struggle with larger training datasets. However, GRU performs better for SMAPE, so once the magnitudes are considered and normalization is applied, GRU may be a better choice.

In summary, in the case of out-of-the-classroom analysis, GRU performs slightly better than LSTM across most metrics (RMSE, MAE, SMAPE) and training ratios, making it the preferred model for this dataset. However, the dataset size for out-of-the-classroom activities is low, and thus, the results in this section are only illustrative. More analyses are needed, with additional features or alternative modeling approaches outside of the scope of our discussion. This, however, does not affect the validity and reliability of the analysis. Furthermore, the inside classroom analysis includes 287 entries across more than 25 time series, and the observations provided can be insightful for the ASEE community.

B List of Acronyms

Acronym	Description
General and Education:	
UDL	Universal Design for Learning
SWD	Students with Disabilities
LMS	Learning Management System
Error Measures:	
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
SMAPE	Symmetric Mean Absolute Percentage Error
Statistical Models:	
ARIMA	Auto-Regressive Integrated Moving Average
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
Machine Learning Models:	
XGBoost	Extreme Gradient Boosting
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
PELT	Pruned Exact Linear Time

Table 7: Acronyms Used in This Article