Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 885–893
https://doi.org/10.22364/bjmc.2025.13.4.07

Predicting Student Performance on a Novel Moodle Dataset Using GRU Time Series Model

Martins SNEIDERS[1], Evalds URTANS[2], Amjed ABU SAA[3]

[1] Riga Technical University, Riga, Latvia
[2] Ventspils University of Applied Sciences, Ventspils, Latvia
[3] Ajman University, UAE

martins.sneiders@gmail.com, evalds.urtans@venta.lv, amjed.abusaa@ku.ac.ae

ORCiD 0009-0006-7961-0822, ORCID 0000-0001-9813-0548, ORCID 0009-0009-4765-6107

Abstract. The prediction of student grades and behavior on online learning platforms is essential for enhancing teaching and learning outcomes. Traditional methods frequently neglect the temporal dynamics inherent in student activity. This study presents a novel Moodle dataset, available on Kaggle, comprising over one million student activity logs. It aims to develop a machine learning time series model designed to predict student grades and behavior, specifically within the context of small universities characterized by limited student populations and course sizes. Moreover, the research seeks to construct a highly accurate predictive model, achieving an accuracy of 83% for semester-long courses. The investigation systematically evaluates the impact of various factors—including course type, passing grade, predictive features, time-splitting schemes, supervised learning algorithms, and performance metrics—on prediction accuracy. The results indicate that GRU (gated recurrent units) models are particularly effective in forecasting student performance in courses that feature active and continuous student engagement. These findings may assist higher education institutions in identifying student performance determinants and establishing an early warning system for potential academic challenges.

Keywords: machine learning, time series, gru, moodle

1 Introduction

Recent advances in machine learning have opened new avenues for analyzing educational data, with time series models emerging as a promising approach for understanding student behavior over time. This study explores the application of a machine learning time series model, specifically one based on gated recurrent units (GRU) to predict student performance using data from Moodle, one of the most popular open-source learning management systems (LMS) worldwide.

A key motivation for this work is the challenge posed by traditional prediction methods, which often fail to capture the dynamic and sequential nature of student interactions. In educational settings, especially in small universities where class sizes are limited, data tend to accumulate slowly, and the temporal patterns of student activity can be critical for early intervention. Moreover, while many institutions have started to implement machine learning models within their LMS, the reliance on black-box approaches raises concerns about prediction reliability and fairness—issues that must be addressed to support sound pedagogical decision-making.

Moodle provides a rich environment for this investigation. Its widespread use and integrated analytics functionalities make it an ideal platform for developing and testing predictive models. Although recent Moodle releases (from version 3.4 onward) support in-platform ML (Machine Learning) model creation, few studies rigorously evaluate these approaches as shown in Tagharobi and Simbeck (2022) and Bognár, Fauszt and Váraljai (2021). This paper not only examines the effectiveness of a GRU-based time series model in forecasting student performance but also discusses the broader implications of integrating machine learning within educational systems. Consistent with prior findings in Daugule et al. (2022), slowly accumulating semester-level data in small cohorts can limit model reliability and timeliness.

Moodle is one of the most popular open-source Learning Management Systems (LMS) in the world with millions of users. Although since the release of Moodle 3.4, it is possible to create ML models within the LMS system, very few studies have been published. Using these models as black boxes poses serious risks of getting unreliable predictions and false alarms as shown in Bognár, Fauszt and Nagy (2021).

Online learning platforms implement ML in their learning analytics functionalities, offering educators predictions on student progress and allowing early interventions. However, these predictive systems are prone to the same fairness problems. If educators use the prediction results in their approach or grading decisions, this can have a major impact on student success. Therefore, educational institutions must assess the fairness of learning analytics before implementing it like shown in Tagharobi and Simbeck (2022).

This work contributes to the growing body of research on educational data analytics by demonstrating how time series models can leverage the rich, sequential data available in LMS platforms like Moodle to provide early warnings and support interventions that enhance student success.

2 Related work

A range of classical machine learning studies has been conducted using engineered features for study course performance prediction. Research utilizing interaction logs, attendance records, quiz submissions, and demographic data demonstrates that when features are meticulously chosen and validated, tree-based models and linear classifiers prove effective in previous work done by Evangelista (2021), Bhutto et al. (2020), Duch et al. (2024). These studies highlight the importance of data quality, the significance of feature importance, and the advantages offered by multi-source attributes. Across institutional datasets, Random Forest often emerges as a strong baseline with key predictors spanning demographics, prior performance, course/instructor attributes, and general factors as shown in Saa et al. (2019). Early-risk benchmarking at 6–60% of course progress reports Random Forest as strongest among algorithms while emphasizing the trade-off between timeliness and accuracy when combining institutional and Moodle logs achieving 84.32% at 20% of course progress and 91.78% at 60% of course progress as listed in Tamada et al. (2021). Also, analysis of Moodle activity confirms a strong link between online engagement patterns and academic achievement, guiding compact and informative feature sets as shown in Shrestha and Pokharel (2021).

Models designed for sequential analysis of temporal behavior have been studied extensively. To effectively capture the sequence and timing of events, RNN-based frameworks such as GRU/LSTM, along with hybrid models like CNN-RNN, have been investigated. These often outperform non-sequential baselines in works like He et al. (2020), Baniata et al. (2024), Aljaloud et al. (2022), Abbaspour et al. (2020), Yin et al. (2023). These approaches can leverage detailed temporal patterns but must be applied with attention to issues of sparsity and the length of the sequences.

Alternative models focus on aspects such as fairness, dependability, and data sparsity as demonstrated in Bognár, Fauszt and Nagy (2021). For instance, a study investigates the critical factors for generating reliable predictions within Moodle, accentuating elements like the size of the predictor matrix, temporal divisions, and evaluation metrics, alongside an examination of the bias–variance dilemma. The issue of data sparsity is particularly pronounced in smaller institutions. It is demonstrated that predictability is significantly influenced by course structure and assessment designs, with self-assessments and quizzes acting as more robust indicators, as evidenced in Kaensar and Wongnin (2023). At the course level, analyses of Moodle indicators—cognitive depth and social breadth—reveal that file/URL interactions constitute weak signals, while self-assessment tests enhance prediction accuracy, and the calculation of quiz indicators proves challenging, as illustrated in Fauszt et al. (2021).

In contrast to previous research, this study presents two main contributions: the release of a new Moodle dataset with over one million logs tailored for small universities, and the benchmarking of a GRU-based temporal model across fine-grained and grouped-grade configurations to tackle scarce long-tail labels. Citations are consistently formatted in author-year style without article titles.

3 Methodology

The main contribution of this research is the Moodle dataset that was obtained from a University of Latvia, enabling the creation of a novel dataset as described in this paper. In accordance with the confidentiality agreement, the university's Moodle data—installed on the institution's server—was provided to the author. Due to institutional constraints and data availability, obtaining additional semesters of data for this analysis was impossible. As a result, the models developed here are based solely on activity and assessment data from one semester. Despite this limitation, the current dataset offers valuable insights into student engagement and performance within the observed semester, and the published data can serve as a benchmark for future research.

After a review of Moodle documentation and publications, the following database tables were selected for use:

1. `mdl_logstore_standard_log` – contains records of various user activities, such as system logins, chapter accesses, answer submissions, and the posting of assessments to students.
2. `mdl_grade_grades` – stores data related to student grades.
3. `mdl_grade_items` – establishes the link between grades and corresponding course information.

Subsequent to data processing, the dataset was published as CSV files:

1. Comprehensive grade data is stored in the file `udk_moodle_all_grades.csv`.
2. Log data is stored in the file `udk_moodle_log.csv`[4].
3. These files are publicly available in the repository, with associated statistics presented in Table 1.

Table 1: Overview of dataset

Metric	udk_moodle_all_grades.csv	udk_moodle_log.csv
Samples	20317	1259411
Columns	id, timemodified, userid, courseid, finalgrade, itemtype	id, timecreated, eventname, action, target, userid, courseid, other
Labels	‘itemtype‘: ("category", "course", "manual", "mod"), ‘userid‘, ‘courseid‘	‘eventname‘, ‘action‘, ‘target‘, ‘userid‘, ‘courseid‘

Analyzing the distribution of grades shown in Table 2, it can be estimated that the majority of students received a "pass" grade, but there are 287 failed grades. A grade of "-1" means that the student did not receive a grade in this subject, and there are 761 such entries.

Table 2: Distribution of grades in dataset

Grade	-1	0	1	2	3	4	5	6	7	8	9	10
Number of samples	761	154	53	23	56	97	162	281	662	1500	2098	1143

An analysis of the grade types presented in Table 3 reveals that the grades are grouped by type. Specifically, the "course" type represents the final grade, while the other types correspond to various assessments administered during the semester. Notably, most of these assessments align with the "mod" type, which indicates that they are generated by different Moodle plugins. To evaluate the prevalence of semester grades,

[4] https://www.kaggle.com/datasets/martinssneiders/moodle-grades-and-action-logs

the column "CWG" (courses with grades) was incorporated into the analysis. The results indicate that the majority of courses are conducted without semester assessments in Moodle; rather, only final grades are recorded, with merely approximately 3.5% of the courses incorporating grading during the semester.

Table 3: Distribution of dataset mark types

	category	**course**	**manual**	**mod**	**CWG**
graded	371	6229	118	7078	216
with ungraded	562	6990	151	12614	255

An analysis of the log file data reveals that the majority of entries were generated by less than 5% of the users and were concentrated in only a few courses. This observation indicates that only a limited number of courses are actively utilizing the Moodle system, possibly those related to information technology; however, this conclusion cannot be definitively verified given the constraints imposed by the anonymized data.

The model employs the Gated Recurrent Unit (GRU) architecture to facilitate efficient time-series modeling while simultaneously conserving memory resources, as shown in Gao and Glowacka (2016).

The hyperparameters of the model include a hidden dimension size of 512, a learning rate of 1e-3, a batch size of 16, and a single-layer GRU model. The model is designed to process Moodle logs by feeding sequential activity data represented as features, followed by a final fully connected layer used for classification into 11-grade categories, with input data being padded for variable length sequences. The training involves the Adam optimizer with a loss function that incorporates class weights to address the class imbalance, which is defined as an alpha coefficient scaled by the inverse frequency of each label in the dataset. Three aggregate representations are computed for each sequence—namely, the element-wise maximum, mean, and the final GRU output—whose concatenation forms a 1536-dimensional vector that is fed into a fully connected layer mapping to 11 output classes, after which a softmax activation is applied to yield a probability distribution over discrete grade categories.

4 Results

An analysis of the dataset facilitated the accurate prediction of student grades, although certain limitations were identified. Specifically, these predictions were achievable solely for a subset of the accessible data. Further scrutiny disclosed that the Moodle platform is actively utilized by fewer than 5% of courses; thus, reliable grade predictions can only be realized for courses that employ continuous grading throughout the semester. This constraint introduces selection bias and restricts the external validity of our findings to Moodle-enabled courses with continuous assessment. Furthermore, the grade distribution is imbalanced, with several grade values being sparsely represented, as observed in Figure 1 and Figure 2. The test dataset was created using a fixed random seed and 20% random sampling without overlapping time sequences.

In the initial iteration of the model, interim assessments conducted during the semester were excluded based on the presumption that they were infrequently utilized by teaching staff, leading to a model founded solely on student activities. However, this approach proved inadequate. Consequently, intermediate evaluation outcomes were subsequently incorporated, and the model was trained exclusively on courses where such evaluations occur. Although this enhances internal consistency, it further restricts the target population to courses with regular intermediate evaluations. Final scores were not analyzed independently but were computed utilizing the highest, lowest, and average values. Preliminary tests revealed that analyses predicated on the highest grade did not yield significant insights. Consequently, emphasis was redirected towards the lowest grade, thereby enabling the prediction of failing students. This decision is supported by the available data, which indicates a predominance of high grades (beginning from 8) in conjunction with a sufficient number of low grades. Accordingly, only those students and courses for which assessments were conducted regularly throughout the instructional process were included in the forecast. Consequently, courses in which students merely accessed reading materials and instructors solely recorded final results were excluded. The forecast input data could then be configured to include or exclude intermediate results as necessary.

As illustrated by a histogram Figure 1, the model demonstrated minor deviations between the predicted and actual marks, culminating in an accuracy of 83%. Nevertheless, this aggregate accuracy warrants cautious interpretation due to class imbalance and low counts per grade. Metrics specific to each grade level, along with ordinal error, offer a more robust evaluation of the model's goodness-of-fit compared to accuracy alone in this context.

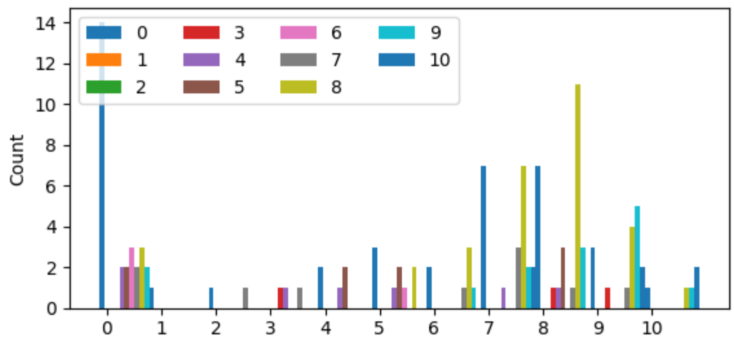

Fig. 1: Histogram of predicted grades indicated by colors and ground truth grades indicated by the X-axis

Additionally, experiments were conducted using the same dataset while excluding grade information and relying solely on student activities. These experiments achieved an accuracy of 80% in predicting final student outcomes, as demonstrated by a hitogram in Figure 2. This result suggests that although intermediate assessment data improves

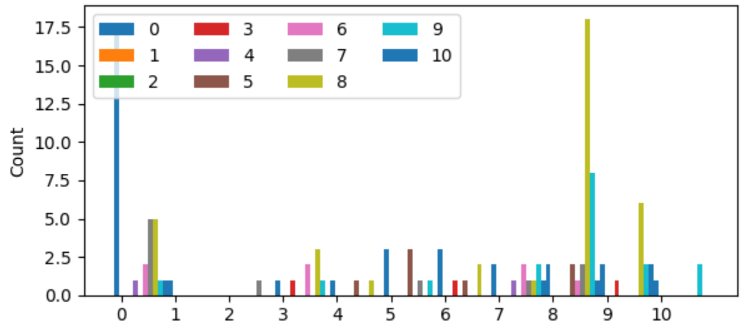

Fig. 2: Histogram of predicted grades indicated by colors and ground truth grades indicated by the X-axis excluding intermediate assessment data

predictive accuracy, a reasonably reliable prediction can still be obtained based solely on user activity data.

5 Further research

In light of the findings of this study, several avenues for future work are evident. One promising area of exploration is the investigation of alternative machine learning models, specifically diverse implementations of Transformer models. Beyond treating grades as an 11-class softmax problem, future work should reformulate the task using ordinal methods—such as cumulative link models or CORN—or as direct regression on $[0, 10]$ with post-hoc rounding. Given skewed and sparsely populated categories, it is also recommended to construct the model with aggregated grade groups (e.g., 0; 1–2; 3–4; 5–6; 7–8; 9–10) and compare performance against the ordinal/regression formulations.

Given that the Moodle platform is presently deployed in less than 5% of courses, it is imperative for future research to focus on enhancing its integration within instructional workflows and evaluating how enriched platform signals impact the modeling capacity and accuracy. To fortify external validity and assess domain shift, it is recommended to broaden the dataset to encompass additional semesters and at least one supplementary institution, utilizing methodologies such as train-on-A/test-on-B and train-on-previous-semester/test-on-next-semester. Privacy-preserving collaboration among universities should be investigated through federated learning to prevent raw data exposure while potentially enhancing generalization.

Evaluation should extend beyond accuracy to include error, agreement, and calibration: report mean absolute error (MAE) of grade points, Quadratic Weighted Kappa, and calibration metrics such as the Brier score and expected calibration error (ECE). To address class imbalance, compare cost-sensitive objectives, focal loss, and class-balanced sampling. Finally, to mitigate sparsity at the course level, investigate transfer and meta learning on "dense" courses and fine-tuning on "sparse" ones—as well as semi-supervised approaches to leverage unlabeled or partially labeled data.

6 Conclusions

This study demonstrates that a compact GRU-based time–series model can forecast student outcomes on a real Moodle deployment from a small university with competitive accuracy, provided that courses exhibit regular, logged engagement and intermediate evaluations. Using a new public dataset comprising 1,259,411 interaction logs and 20,317 grade records. In terms of coverage and adoption, only a small fraction of courses systematically use Moodle for continuous assessment, with 6,990 course-level (final) grade records observed, of which 6,229 are graded and 761 are ungraded. Courses with regular semester assessments account for approximately 3.5% relative to the recorded course finals, and usage is highly skewed with fewer than 5% of users and a handful of courses generating the majority of logs. Regarding the outcome distribution and imbalance, among the 6,229 graded course finals, high marks in the range of 8–10 make up 76.1%, while low marks between 0–3 are rare at 4.6%. Depending on the pass threshold set by institutions, the proportion of fails varies, highlighting the relevance of the long-tailed label distribution for both class weighting and ordinal-aware evaluation. In terms of predictive performance and error shape, including intermediate assessment signals enables the GRU model to achieve 83% accuracy on the test set, with activity logs alone leading to 80% accuracy. errors tend to be small and ordinally local, with a high proportion of predictions falling within one grade point of the true grade, though very low grades suffer from noisier estimates due to limited test samples, which emphasizes the necessity for imbalance-aware training and reporting. Despite sparse adoption and skewed outcomes paired with careful scope limitations, calibration, and ordinal-aware evaluation, such models can underpin practical, low-latency early-warning systems for universities.

References

Abbaspour, S., Fotouhi, F., Sedaghatbaf, A., Fotouhi, H., Vahabi, M., Lindén, M. (2020). A comparative analysis of hybrid deep learning models for human activity recognition, *Sensors (Basel, Switzerland)* **20**.
`https://api.semanticscholar.org/CorpusID:222255551`

Aljaloud, A. S., Uliyan, D. M., Alkhalil, A., Elrhman, M. A., Alogali, A. F. M., Altameemi, Y. M., Altamimi., M., Kwan, P. (2022). A deep learning model to predict student learning outcomes in lms using cnn and lstm, *IEEE Access* **10**, 85255–85265.
`https://api.semanticscholar.org/CorpusID:251378155`

Baniata, L. H., Kang, S., Alsharaiah, M., Baniata, M. H. (2024). Advanced deep learning model for predicting the academic performances of students in educational institutions, *Applied Sciences* .
`https://api.semanticscholar.org/CorpusID:268136123`

Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., Anwar, M. (2020). Predicting students' academic performance through supervised machine learning, *2020 International Conference on Information Science and Communication Technology (ICISCT)* pp. 1–6.
`https://api.semanticscholar.org/CorpusID:218468275`

Bognár, L., Fauszt, T., Nagy, G. (2021). Analysis of conditions for reliable predictions by moodle machine learning models, *Int. J. Emerg. Technol. Learn.* **16**.

Bognár, L., Fauszt, T., Váraljai, M. (2021). The impact of online quizzes on student success, *Int. J. Emerg. Technol. Learn.* **16**.

Daugule, I., Kapenieks, A., Timsans, Z. (2022). Use of knowledge acquisition surface to monitor and assess students' success, *Int. J. Emerg. Technol. Learn.* **17**, 109–125.
`https://api.semanticscholar.org/CorpusID:251124300`

Duch, D., May, M., George, S. (2024). Enhancing predictive analytics for students' performance in moodle: Insight from an empirical study, *Journal of Data Science and Intelligent Systems* .
`https://api.semanticscholar.org/CorpusID:272935088`

Evangelista, E. D. L. (2021). A hybrid machine learning framework for predicting students' performance in virtual learning environment, *Int. J. Emerg. Technol. Learn.* **16**.

Fauszt, T., Bognár, L., Sándor, Á. (2021). Increasing the prediction power of moodle machine learning models with self-defined indicators, *Int. J. Emerg. Technol. Learn.* **16**.

Gao, Y., Glowacka, D. (2016). Deep gate recurrent neural network, *Proceedings of the Asian Conference on Machine Learning (ACML)*, Vol. 63, pp. 1–16.

He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., Jiang, B. (2020). Online at-risk student identification using rnn-gru joint neural networks, *Inf.* **11**, 474.
`https://api.semanticscholar.org/CorpusID:225166054`

Kaensar, C., Wongnin, W. (2023). Analysis and prediction of student performance based on moodle log data using machine learning techniques, *Int. J. Emerg. Technol. Learn.* **18**, 184–203.
`https://api.semanticscholar.org/CorpusID:258881108`

Saa, A. A., Al-Emran, M., Shaalan, K. F. (2019). Mining student information system records to predict students' academic performance, *International Conference on Advanced Machine Learning Technologies and Applications*.

Shrestha, S., Pokharel, M. (2021). Educational data mining in moodle data, *International Journal of Informatics and Communication Technology* **10**, 9–18.

Tagharobi, H., Simbeck, K. (2022). Introducing a framework for code based fairness audits of learning analytics systems on the example of moodle learning analytics, *International Conference on Computer Supported Education*.

Tamada, M. M., Giusti, R., de Magalhães Netto, J. F. (2021). Predicting student performance based on logs in moodle lms, *2021 IEEE Frontiers in Education Conference (FIE)* pp. 1–8.
`https://api.semanticscholar.org/CorpusID:245422775`

Yin, C., Tang, D., Zhang, F., Tang, Q., Feng, Y., He, Z. (2023). Students learning performance prediction based on feature extraction algorithm and attention-based bidirectional gated recurrent unit network, *PLOS ONE* **18**.
`https://api.semanticscholar.org/CorpusID:264487735`

Received September 7, 2023 , revised September 15, 2025, accepted November 13, 2025