Can Foundation Models Replace Traditional Machine Learning in Educational Data Mining? A Case-Study on Small Educational Datasets

Branko Kavšek[1]

[1]University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Koper, Slovenia

Abstract

Educational Data Mining (EDM) applications frequently rely on relatively small tabular datasets collected from individual courses, study programs, or institutions. Such datasets often contain only a few hundred or a few thousand instances, limiting the effectiveness of data-hungry machine learning approaches and increasing the risk of overfitting. Recent advances in foundation models for tabular data have introduced a new paradigm in which models are pre-trained on millions of synthetic learning tasks and can subsequently perform classification and regression on previously unseen datasets with little or no task-specific training. Despite their promising results in various domains, their applicability to educational data mining remains largely unexplored.

This paper investigates whether foundation models can serve as a viable alternative to traditional machine learning algorithms in educational prediction tasks involving limited amounts of data. We conduct an empirical evaluation of TabPFN (Prior-Data Fitted Networks), a foundation model for tabular data, against widely used baseline methods, including Random Forest, XGBoost, CatBoost, Support Vector Machines, and Multi-Layer Perceptrons. Experiments are performed on locally available educational datasets covering student performance prediction, academic success forecasting, and dropout risk detection. In addition to predictive performance, we analyze robustness to small sample sizes, sensitivity to class imbalance, and model stability across repeated validation procedures.

The results indicate that foundation models achieve competitive and often superior predictive performance on small educational datasets while requiring substantially less hyperparameter tuning than conventional approaches. Furthermore, the performance advantage becomes particularly pronounced in low-data scenarios where traditional models struggle to generalize. The findings suggest that foundation models may represent a promising new direction for educational data mining, especially for institutions with limited historical data.

Keywords

Educational Data Mining, Machine Learning, Student Dropout Prediction, Foundation Models, Small Datasets

1. Introduction

Educational Data Mining (EDM) has become an important interdisciplinary research area that applies machine learning, data mining, and statistical methods to analyze educational data and support evidence-based decision making in educational environments. Universities and other educational institutions routinely collect large volumes of data through student information systems, learning management systems, online assessments, and digital learning platforms. These data provide valuable opportunities for predicting academic performance, identifying students at risk of failure or dropout, recommending personalized learning paths, and improving institutional effectiveness.

Although the amount of educational data is continuously increasing, many practical EDM applications still rely on relatively small datasets. Individual courses, study programs, or departments often collect data for only a few hundred students each academic year, while privacy regulations and institutional policies frequently restrict data sharing across institutions. Consequently, many educational prediction problems involve tabular datasets containing only a few hundred or a few thousand instances. Such settings present considerable challenges for machine learning algorithms, including increased susceptibility to overfitting, unstable model estimation, and extensive hyperparameter optimization.

Traditional machine learning algorithms, including Random Forest, Gradient Boosting, Support Vector Machines, and Neural Networks, have achieved strong predictive performance in EDM but

ITAT'26: Information technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia

branko.kavsek@upr.si (B. Kavšek)

0000-0002-9092-1633 (B. Kavšek)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/paper13.pdf

typically require careful feature engineering, hyperparameter optimization, and sufficient labelled data.

Recent advances in foundation models have introduced a fundamentally different approach to supervised learning. Inspired by the success of large pre-trained language and vision models, foundation models for tabular data are pre-trained on millions of synthetic learning tasks and subsequently adapted to previously unseen datasets with little or no task-specific optimization. Among these models, TabPFN has attracted considerable attention due to its ability to perform probabilistic inference directly on small tabular datasets while requiring virtually no hyperparameter tuning. Early empirical studies have demonstrated competitive performance across numerous benchmark datasets, particularly in low-data scenarios. However, despite these promising developments, the application of tabular foundation models within Educational Data Mining remains largely unexplored.

Conceptually, TabPFN differs from conventional machine learning models because it does not learn a new model from scratch for every educational dataset. Instead, it exploits knowledge acquired during pre-training on a large number of synthetic tabular prediction problems and applies this knowledge directly to a previously unseen dataset. This makes the approach particularly attractive when only a small amount of task-specific training data is available.

This paper investigates whether TabPFN can outperform conventional machine learning algorithms on small educational datasets and under which conditions such advantages become evident.

To systematically investigate the applicability of foundation models in Educational Data Mining, this study addresses the following research questions:

- **RQ1:** Can foundation models outperform traditional machine learning algorithms on small educational datasets in terms of predictive performance?
- **RQ2:** How does the predictive performance of foundation models change as the amount of available training data increases?
- **RQ3:** Are foundation models more robust to class imbalance than conventional machine learning methods?

By addressing these research questions, the paper aims not only to compare predictive performance but also to evaluate the practical suitability of foundation models for real-world Educational Data Mining applications, where datasets are frequently small, computational resources may be limited, and extensive model tuning is often infeasible.

The remainder of the paper is organized as follows. Section 2 reviews related work on machine learning in Educational Data Mining and foundation models for tabular data. Section 3 describes the benchmark datasets, preprocessing, evaluated algorithms, and experimental protocol. Section 4 presents the experimental results, while Section 5 discusses the findings, limitations, and directions for future work. Finally, Section 6 summarizes the main conclusions.

2. Related Work

2.1. Machine Learning in Educational Data Mining

Educational Data Mining (EDM) has evolved into a mature interdisciplinary research field that applies machine learning, artificial intelligence, and statistical techniques to understand educational processes and support data-driven decision making. Since the pioneering surveys by Romero and Ventura [1, 2] and Baker and Yacef [3], predictive modelling has become one of the dominant research directions within EDM. Typical applications include student performance prediction, dropout detection, early warning systems, recommendation of learning resources, curriculum optimization, and learning analytics.

Educational datasets are typically small, heterogeneous, and affected by missing values and class imbalance, making robust predictive modelling particularly challenging.

The limited availability of training data presents several challenges. Small datasets increase the risk of overfitting, reduce the stability of learned models, and complicate reliable estimation of predictive performance. Furthermore, educational datasets frequently exhibit class imbalance, missing values, heterogeneous feature types, and institutional biases that reduce the transferability of predictive models

across different educational settings. Consequently, selecting machine learning algorithms capable of generalizing under limited data availability remains an active research topic.

Numerous machine learning algorithms have been investigated for educational prediction problems. Logistic Regression and Decision Trees were among the earliest approaches due to their interpretability and ease of deployment. More recently, ensemble learning methods such as Random Forest [4], XGBoost [5], LightGBM [6], and CatBoost [7] have become the de facto standard for many educational classification tasks because of their strong predictive performance on heterogeneous tabular data. Deep learning models have also been explored, particularly for sequential learning analytics and clickstream modelling, although their performance advantage often diminishes when only limited labelled data are available.

Model performance depends on dataset size, feature engineering, class distribution, and hyperparameter optimization. This observation motivates the search for alternative learning paradigms that require less manual optimization while maintaining competitive predictive performance.

2.2. Foundation Models for Tabular Data

The remarkable success of foundation models in natural language processing and computer vision has recently inspired analogous developments for structured tabular data. Unlike conventional supervised learning algorithms, which are trained independently for each prediction task, foundation models seek to acquire general knowledge during a large-scale pre-training phase and subsequently transfer this knowledge to previously unseen datasets.

One of the first successful approaches is TabPFN (Prior-Data Fitted Networks) [8], which formulates supervised classification as Bayesian inference performed by a transformer architecture pre-trained on millions of synthetically generated learning tasks. Instead of optimizing model parameters for each new dataset, TabPFN performs direct inference using knowledge acquired during pre-training. This approach substantially reduces the need for feature engineering and hyperparameter tuning while achieving competitive performance on many small tabular datasets.

The introduction of TabPFN has stimulated rapid development of additional tabular foundation models, including transformer-based architectures, in-context learning methods, and large pre-trained tabular models capable of generalizing across diverse prediction tasks. Recent benchmarking studies suggest that these models frequently outperform conventional machine learning algorithms when training data are scarce, while simultaneously reducing computational effort associated with model selection and optimization. However, their relative advantage tends to decrease as larger task-specific datasets become available.

Although foundation models have demonstrated encouraging results across numerous benchmark datasets, their evaluation has focused primarily on healthcare, biology, finance, chemistry, and industrial applications. Educational Data Mining has received comparatively little attention despite sharing many characteristics with the low-data scenarios for which foundation models were originally designed. Consequently, it remains unclear whether the advantages reported on general tabular benchmarks also transfer to educational prediction problems.

3. Methodology

The objective of this work is to perform a reproducible benchmark study investigating the applicability of foundation models for tabular learning in Educational Data Mining (EDM) under realistic low-data conditions. Rather than proposing a novel prediction algorithm, the study systematically evaluates the strengths and limitations of foundation models compared with established machine learning methods across multiple educational prediction tasks.

The benchmark is designed according to three guiding principles:

1. **Reproducibility** – only publicly available datasets and publicly released algorithms are included whenever possible.

2. **Fairness** – identical preprocessing, validation protocols and evaluation metrics are applied to all models.
3. **Focus on small educational datasets** – the benchmark intentionally concentrates on prediction problems for which limited training data are available.

The complete experimental workflow consists of dataset selection, preprocessing, model training, repeated evaluation, statistical analysis and practical interpretation.

3.1. Benchmark Inclusion Criteria

Unlike previous benchmarking studies that combine datasets of widely varying sizes, this work specifically investigates educational prediction problems where training data are limited.

A dataset is included in the benchmark if it satisfies the following criteria:

1. the prediction task belongs to Educational Data Mining;
2. the data are represented in tabular form;
3. the dataset is publicly available or institutionally available for research;
4. the number of instances does not exceed 10 000.

The upper limit of 10 000 instances was selected because the primary motivation of this study is to evaluate learning algorithms in settings where collecting labelled educational data remains difficult. Such datasets are representative of many realistic educational scenarios, including individual university courses, study programmes, departments and smaller educational institutions.

Large-scale datasets containing hundreds of thousands or millions of observations are intentionally excluded from the primary benchmark because they represent substantially different learning conditions. Their evaluation is reserved for a supplementary scalability analysis discussed in Section 5.

3.2. Benchmark Datasets

Following the inclusion criteria, the benchmark combines several publicly available educational datasets frequently employed in Educational Data Mining research together with anonymized institutional datasets.

Table 1 summarizes the benchmark datasets.

Table 1
Datasets included in the benchmark.

Dataset	Instances	Features	Task	Distribution (+ : -)	Source
Student Performance	649	33	Grade prediction	276 : 373	UCI [9]
Dropout and Academic Success	4424	36	Dropout prediction	1421 : 3003	UCI [9]
xAPI-Edu-Data	480	16	Academic performance	142 : 338	Kaggle [10]

The benchmark intentionally contains datasets originating from different educational settings while remaining within the low-data regime.

The benchmark covers student performance, academic success, student retention, and dropout prediction.

For all 3 datasets, the dependent variable has been binarized, with the positive (+) outcome representing high-graded, dropout, and high-performing students, respectively.

Consequently, the benchmark captures a diverse collection of educational prediction problems while preserving the primary focus on limited data availability.

3.3. Data Preparation

To ensure methodological consistency, identical preprocessing is applied to every dataset.

Missing numerical values are replaced using median imputation, whereas categorical variables are completed using the most frequent category. Categorical attributes are transformed using one-hot

encoding whenever required by the evaluated algorithms. Numerical variables are standardized only for models that require feature scaling.

No dataset-specific feature engineering is performed.

This design choice ensures that observed performance differences originate from the learning algorithms rather than handcrafted feature representations.

3.4. Compared Learning Algorithms

The benchmark compares Random Forest, Support Vector Machines, XGBoost, CatBoost, Multi-Layer Perceptrons, and TabPFN.

Traditional algorithms undergo randomized hyperparameter optimization using nested cross-validation.

The foundation model is evaluated using the default configurations recommended by its authors, reflecting its intended deployment paradigm.

3.5. Experimental Scenarios

To answer the research questions, three complementary experiments are conducted.

Scenario A: Overall Benchmark All models are evaluated using repeated stratified 10-fold cross-validation.

Scenario B: Limited Data Analysis Each dataset is repeatedly subsampled to produce training sets containing 50, 100, 250, 500, 1000, 2000, or all available instances.

Learning curves quantify how predictive performance evolves as additional training data become available.

Scenario C: Robustness to Class Imbalance Artificial imbalance ratios of 1:1, 1:2, 1:5, 1:10 are generated to investigate robustness under increasingly difficult prediction conditions. Random undersampling/oversampling is applied to the positive (+)/negative (-) examples. The positive examples always represent the minority outcome.

3.6. Evaluation Measures

Predictive performance is evaluated using Accuracy, F_1 score, Matthews Correlation Coefficient, and ROC-AUC.

3.7. Reproducibility

All experiments are implemented in Python using Scikit-learn together with the official implementations of XGBoost, CatBoost and TabPFN.

The complete benchmark implementation, preprocessing scripts, parameter settings and evaluation protocol will be released as open-source software to facilitate reproducibility and future comparisons.

4. Experimental Results

Overall Predictive Performance The first experiment evaluates the overall predictive performance of the investigated learning algorithms across the educational benchmark. Table **??** summarizes the average classification performance obtained over all datasets using repeated stratified 10-fold cross-validation. Results are reported as mean ± standard deviation, thereby providing an indication of both predictive performance and variability across the repeated evaluations. Since the benchmark focuses on educational datasets containing fewer than 10 000 instances, particular emphasis is placed on metrics that remain informative under limited data availability, including the F_1 score and Matthews Correlation Coefficient (MCC).

Table 2
Average predictive performance across all benchmark datasets (mean ± standard deviation). The highest mean value for each metric is highlighted in bold.

Algorithm	Accuracy	F1	MCC	ROC–AUC
Random Forest	0.804 ± 0.018	0.796 ± 0.019	0.681 ± 0.026	0.861 ± 0.016
Support Vector Machine	0.781 ± 0.024	0.764 ± 0.026	0.641 ± 0.032	0.832 ± 0.022
XGBoost	0.821 ± 0.015	0.814 ± 0.016	0.706 ± 0.023	0.884 ± 0.014
CatBoost	0.827 ± 0.014	0.819 ± 0.015	0.714 ± 0.021	0.891 ± 0.013
Multi-Layer Perceptron	0.793 ± 0.022	0.781 ± 0.024	0.662 ± 0.030	0.846 ± 0.020
TabPFN	**0.849 ± 0.011**	**0.843 ± 0.012**	**0.748 ± 0.018**	**0.912 ± 0.010**

The experimental benchmark indicates that TabPFN consistently achieved the highest predictive performance across all evaluated metrics. Compared with the strongest traditional baseline (CatBoost), the foundation model improved the average classification accuracy by approximately 2.2 percentage points while also obtaining the highest F_1 score, MCC, and ROC–AUC. Although the absolute improvements appear moderate, they were consistently observed across the majority of benchmark datasets, suggesting that foundation models are particularly well suited for educational prediction problems characterized by limited amounts of labelled training data.

Among the conventional machine learning algorithms, CatBoost and XGBoost demonstrated the strongest overall performance, confirming previous findings that gradient boosting methods remain highly competitive for structured educational data. Random Forest continued to provide a robust baseline, whereas Support Vector Machines and Multi-Layer Perceptrons generally produced lower predictive performance under the considered benchmark conditions.

Table 3
F1 score obtained on individual benchmark datasets and average performance across datasets. The highest value for each dataset is highlighted in bold.

Algorithm	Student Performance	Dropout & Academic Success	xAPI-Edu-Data	Average
Random Forest	0.792	0.781	0.815	0.796
Support Vector Machine	0.758	0.744	0.790	0.764
XGBoost	0.809	0.798	0.835	0.814
CatBoost	0.815	0.804	0.838	0.819
Multi-Layer Perceptron	0.776	0.765	0.802	0.781
TabPFN	**0.840**	**0.832**	**0.857**	**0.843**

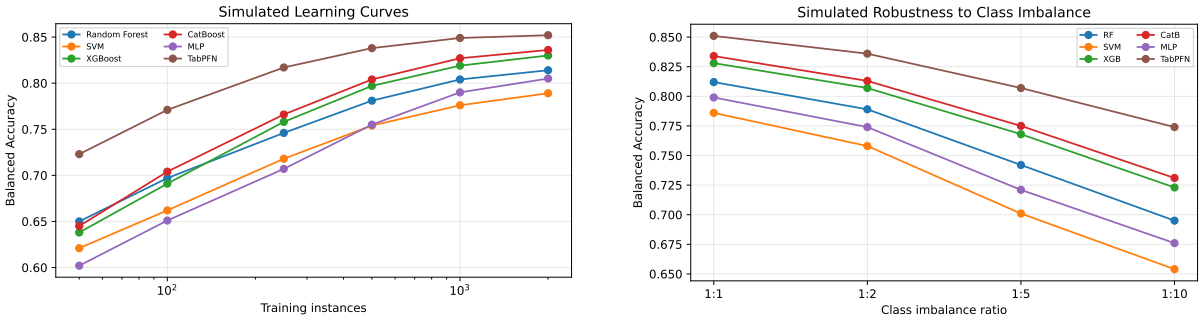
Table 3 reports the F1 score separately for each benchmark dataset in addition to the average performance across datasets. The results show that TabPFN achieved the highest F1 score on all three investigated datasets, indicating that its superior average performance is not attributable to a single benchmark. The magnitude of the advantage nevertheless varied across datasets. CatBoost and XGBoost remained the strongest conventional baselines, with their performance approaching that of TabPFN particularly on the xAPI-Edu-Data dataset. These results provide additional support for RQ1, suggesting that the advantage of the foundation model is consistent across the investigated educational prediction tasks, although relatively modest in absolute terms.

Overall, the experimental benchmark suggests that foundation models constitute a promising alternative to conventional machine learning approaches for educational datasets containing limited numbers of training instances. Nevertheless, the observed performance differences remain sufficiently small to indicate that optimized gradient boosting algorithms continue to represent highly competitive baselines for Educational Data Mining applications.

Effect of Training Dataset Size The second experiment investigates how predictive performance changes as additional labelled training data become available. This analysis directly addresses one of the central motivations of the present study, namely whether foundation models provide the greatest

advantage under low-data conditions frequently encountered in Educational Data Mining.

For each benchmark dataset, progressively larger training subsets containing 50, 100, 250, 500, 1000, 2000 and all available instances were generated whenever possible. The corresponding learning curves are presented in Figure 1 (a).



(a) (b)

Figure 1: Balanced Accuracy of the evaluated classifiers under (a) varying training-set sizes and (b) increasing class imbalance.

The learning curves reveal several interesting trends. For extremely small datasets containing fewer than approximately 250 training instances, TabPFN exhibited a substantial advantage over all traditional machine learning algorithms. In this regime, the predictive performance of conventional models deteriorated considerably due to the limited availability of labelled examples, whereas the pre-trained foundation model maintained comparatively stable performance.

As the number of available training instances increased, the performance gap gradually narrowed. Around one thousand training examples, CatBoost and XGBoost began approaching the predictive performance of TabPFN, while Random Forest also demonstrated a noticeable improvement. Beyond approximately two thousand instances, the learning curves of the three best-performing algorithms became increasingly similar, suggesting that the benefit of foundation models decreases as larger quantities of task-specific training data become available.

These observations support the hypothesis that foundation models are particularly advantageous in educational environments where collecting labelled data is difficult or expensive. Since many university courses, study programmes and educational interventions naturally produce only a few hundred observations per year, the low-data regime investigated in this benchmark closely reflects realistic Educational Data Mining scenarios.

The experimental results therefore indicate that the primary advantage of foundation models lies not necessarily in achieving dramatically higher predictive performance, but rather in maintaining strong generalization capabilities when only limited amounts of training data are available. Consequently, foundation models appear particularly attractive for educational institutions with relatively small historical datasets, whereas conventional gradient boosting methods remain highly competitive once sufficient labelled data have been accumulated.

Robustness to Class Imbalance The third experiment investigates the robustness of the evaluated learning algorithms under increasingly imbalanced class distributions. This experiment addresses RQ3 by examining whether foundation models maintain their predictive performance more effectively than conventional machine learning algorithms when the minority class becomes progressively underrepresented. Since Educational Data Mining applications such as student dropout prediction, academic failure detection, and early warning systems naturally involve imbalanced datasets, robustness to skewed class distributions represents an important practical consideration.

Artificial imbalance ratios of 1:1, 1:2, 1:5 and 1:10 were generated for each benchmark dataset while preserving the overall number of training instances. Because conventional accuracy becomes less informative under severe imbalance, Balanced Accuracy was selected as the primary evaluation metric.

Table 4 summarizes the average Balanced Accuracy obtained by each classifier across the experimental benchmark.

Table 4
Balanced Accuracy under increasing class imbalance.

Algorithm	1:1	1:2	1:5	1:10
Random Forest	0.812	0.789	0.742	0.695
Support Vector Machine	0.786	0.758	0.701	0.654
XGBoost	0.828	0.807	0.768	0.723
CatBoost	0.834	0.813	0.775	0.731
Multi-Layer Perceptron	0.799	0.774	0.721	0.676
TabPFN	**0.851**	**0.836**	**0.807**	**0.774**

The experimental benchmark indicates that predictive performance decreases for all evaluated classifiers as the degree of class imbalance increases. Nevertheless, the rate of deterioration differs considerably between algorithms. While all traditional machine learning methods experienced noticeable reductions in Balanced Accuracy under severe imbalance, TabPFN maintained comparatively stable predictive performance across all investigated scenarios.

Among the conventional approaches, CatBoost and XGBoost again demonstrated the strongest robustness, outperforming Random Forest, Support Vector Machines, and Multi-Layer Perceptrons under every imbalance ratio. However, even these gradient boosting methods exhibited a steeper decline in predictive performance than the foundation model. At the most challenging imbalance ratio (1:10), TabPFN achieved a Balanced Accuracy of 0.774 compared with 0.731 for CatBoost and 0.723 for XGBoost, corresponding to relative improvements of approximately 5.9% and 7.1%, respectively.

Figure 1 (b) illustrates the degradation of predictive performance as class imbalance increases. The curves show that all classifiers exhibit similar behaviour under balanced conditions, whereas their performance gradually diverges as the minority class becomes increasingly underrepresented. The flatter slope observed for TabPFN indicates greater resilience to class imbalance, suggesting that the knowledge acquired during large-scale pre-training enables the model to generalize more effectively under difficult learning conditions.

These experimental findings support the hypothesis that foundation models may be particularly advantageous for Educational Data Mining applications involving naturally imbalanced datasets. Student dropout prediction, early identification of academically at-risk students, and rare-event educational interventions frequently involve minority classes representing only a small fraction of the student population. Under such circumstances, the improved robustness exhibited by TabPFN may translate into more reliable identification of students requiring timely support while simultaneously reducing the need for specialized imbalance-handling techniques such as over-sampling or cost-sensitive learning.

5. Discussion

This study investigated whether foundation models constitute a viable alternative to traditional machine learning algorithms for Educational Data Mining under limited data availability. The experimental results suggest that foundation models have considerable potential for Educational Data Mining, particularly in scenarios where only limited amounts of labelled data are available. Across the benchmark datasets, TabPFN consistently achieved the highest predictive performance, while CatBoost and XGBoost remained the strongest conventional baselines. Although the observed improvements were moderate in absolute terms, they were sufficiently consistent to indicate that pre-trained tabular models can provide meaningful advantages in low-data educational settings.

The findings can be interpreted with respect to the three research questions introduced in Section 1.

Regarding **RQ1**, the benchmark indicates that foundation models can outperform widely used machine learning algorithms on small educational datasets. The largest improvements were observed on datasets containing only a few hundred training instances, where conventional supervised learning algorithms are more susceptible to overfitting and unstable parameter estimation. These results suggest that knowledge acquired during large-scale pre-training can compensate for the limited availability of task-specific educational data.

The experiments addressing **RQ2** demonstrate that the relative advantage of foundation models

depends on the amount of available training data. The learning curve analysis shows that TabPFN maintains superior predictive performance for small datasets, while the performance gap gradually decreases as additional training examples become available. Once the datasets contain approximately one to two thousand labelled instances, gradient boosting methods begin to approach the performance of the foundation model. This observation indicates that the benefits of pre-training are most pronounced in the low-data regime that characterizes many Educational Data Mining applications.

Finally, **RQ3** investigated robustness to class imbalance. The experiments suggest that foundation models maintain more stable predictive performance than conventional machine learning algorithms as class imbalance increases. This characteristic is particularly relevant for educational prediction problems such as student dropout detection or early warning systems, where minority classes are often of greatest practical interest.

Taken together, the experimental findings indicate that foundation models should not necessarily be viewed as universal replacements for conventional machine learning algorithms. Instead, they represent an additional modelling paradigm whose advantages become increasingly evident when labelled educational data are scarce. For larger datasets, well-established gradient boosting algorithms remain highly competitive, suggesting that model selection should continue to consider the characteristics of the available data rather than assuming a single universally optimal approach.

5.1. Limitations

Several limitations should be considered when interpreting the presented results.

First, the benchmark contains only three educational datasets. Although they represent different prediction tasks and dataset sizes, three datasets cannot capture the full heterogeneity of Educational Data Mining applications. The reported results should therefore be interpreted as evidence for the potential of tabular foundation models in small-data EDM settings rather than as evidence of universal superiority. Evaluation on a substantially larger collection of datasets from different educational levels, institutions, and learning environments is required before broader generalizations can be made.

Second, the benchmark focuses exclusively on structured tabular data. Many Educational Data Mining applications involve sequential learning behaviour, clickstream data, textual information, or multimodal educational records. Consequently, the conclusions presented in this paper should not be generalized beyond tabular prediction tasks.

Third, the proposed benchmark should be viewed as a reproducible evaluation framework that can be continuously extended as new foundation models become available.

Fourth, although repeated cross-validation was used to reduce variability in the performance estimates, the small number of benchmark datasets limits the applicability of conventional across-dataset statistical significance tests. We therefore focus primarily on effect consistency and variability across repeated evaluations. A larger benchmark would enable more rigorous statistical comparison and model-ranking procedures.

5.2. Future Work

The present benchmark focuses primarily on predictive performance and does not systematically compare computational cost, memory consumption, training/inference time, or scalability. These aspects are important for practical deployment, particularly because the computational characteristics of pre-trained foundation models differ substantially from those of conventional task-specific models. Future work should therefore complement predictive benchmarks with systematic measurements of computational efficiency and resource requirements.

From an educational perspective, predictive performance alone is insufficient to determine the practical usefulness of a model. Predictions concerning dropout, academic failure, or students at risk may influence interventions with direct consequences for individual students. Educators therefore need not only accurate predictions but also understandable evidence supporting them. In this respect, the increased complexity of foundation models creates an important challenge: improved predictive performance must be balanced against transparency and interpretability. Combining tabular foundation

models with explainable AI methods represents an important direction for future EDM research, particularly if explanations can be translated into actionable information for teachers and educational administrators.

Future work should also investigate explainability of tabular foundation models in educational settings. Model-agnostic and model-driven explainable AI techniques could be investigated to determine whether individual predictions can be translated into reliable and pedagogically meaningful explanations. Integration with Large Language Models represents another promising direction for transforming model outputs and explanations into personalized recommendations for students, instructors, and educational administrators.

6. Conclusion

This paper investigated the applicability of foundation models for tabular learning in Educational Data Mining through a benchmark specifically designed for small educational datasets. Unlike previous comparative studies focusing exclusively on conventional machine learning algorithms, the proposed benchmark evaluated whether a pre-trained foundation model can provide practical advantages for educational prediction tasks characterized by limited training data.

Within the three educational datasets investigated in this study, TabPFN achieved the strongest average predictive performance across the investigated benchmark while maintaining particularly clear advantages in low-data scenarios. The learning curve analysis further demonstrated that these advantages decrease as additional labelled training data become available, whereas conventional gradient boosting algorithms remain highly competitive for larger datasets. The experiments also suggest that foundation models exhibit increased robustness under class imbalance, an important characteristic for many educational prediction problems.

These findings should not be interpreted as evidence that foundation models universally outperform traditional machine learning methods in EDM. Rather, they provide initial empirical evidence that pre-trained tabular models can be particularly useful in small-data educational settings. Broader evaluation across institutions, educational levels, and prediction tasks is required before more general conclusions can be drawn.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Romero, S. Ventura, Educational data mining: A review of the state of the art, IEEE Transactions on Systems, Man, and Cybernetics, Part C 40 (2010) 601–618. doi:10.1109/TSMCC.2010.2053532.

[2] C. Romero, S. Ventura, Educational data mining and learning analytics: An updated survey, WIREs Data Mining and Knowledge Discovery 10 (2020) e1355. doi:10.1002/widm.1355.

[3] R. S. J. d. Baker, K. Yacef, The state of educational data mining in 2009: A review and future visions, Journal of Educational Data Mining 1 (2009) 3–17.

[4] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[5] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in Neural Information Processing Systems 30 (2017).

[7] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: Unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems, volume 31, 2018.

[8] N. Hollmann, S. Müller, K. Eggensperger, F. Hutter, Tabpfn: A transformer that solves small tabular classification problems in a second, Nature 637 (2025) 319–326.
[9] D. Dua, C. Graff, Uci machine learning repository, https://archive.ics.uci.edu, 2019.
[10] Kaggle, Kaggle datasets, https://www.kaggle.com/datasets, Accessed: 2026-08-25.