applied sciences

MDPI

Article

A Comparative Analysis of Student Performance Prediction: Evaluating Optimized Deep Learning Ensembles Against Semi-Supervised Feature Selection-Based Models

Jose Antonio Lagares Rodríguez *, Norberto Díaz-Díaz and Carlos David Barranco González

Department of Computer Science, Intelligent Data Analysis Group (DATAi), Universidad Pablo de Olavide, 41013 Seville, Spain; ndiaz@upo.es (N.D.-D.); cdbargon@upo.es (C.D.B.G.)
* Correspondence: jalagrod@upo.es

Abstract: Advancements in modern technology have significantly increased the availability of educational data, presenting researchers with new challenges in extracting meaningful insights. Educational Data Mining offers analytical methods to support the prediction of student outcomes, development of intelligent tutoring systems, and curriculum optimization. Prior studies have highlighted the potential of semi-supervised approaches that incorporate feature selection to identify factors influencing academic success, particularly for improving model interpretability and predictive performance. Many feature selection methods tend to exclude variables that may not be individually powerful predictors but can collectively provide significant information, thereby constraining a model's capabilities in learning environments. In contrast, Deep Learning (DL) models paired with Automated Machine Learning techniques can decrease the reliance on manual feature engineering, thereby enabling automatic fine-tuning of numerous model configurations. In this study, we propose a reproducible methodology that integrates DL with AutoML to evaluate student performance. We compared the proposed DL methodology to a semi-supervised approach originally introduced by Yu et al. under the same evaluation criteria. Our results indicate that DL-based models can provide a flexible, data-driven approach for examining student outcomes, in addition to preserving the importance of feature selection for interpretability. This proposal is available for replication and additional research.

Keywords: deep learning (DL); machine learning (ML); feature selection (FS); educational data mining (EDM); artificial neuronal networks (ANNs); AutoML; xAPI

Academic Editor: Andrea Prati

Received: 27 January 2025
Revised: 6 April 2025
Accepted: 16 April 2025
Published: 26 April 2025

Citation: Lagares Rodríguez, J.A.; Díaz-Díaz, N.; Barranco González, C.D. A Comparative Analysis of Student Performance Prediction: Evaluating Optimized Deep Learning Ensembles Against Semi-Supervised Feature Selection-Based Models. *Appl. Sci.* **2025**, *15*, 4818. https://doi.org/10.3390/app15094818

1. Introduction

The increasing integration of digital technologies into educational environments has substantially expanded the volume and diversity of data available for analysis, creating a need for advanced analytical techniques capable of revealing meaningful insights. In response, Educational Data Mining (EDM) has emerged as a research area that applies data-driven methods to identify learning patterns, predict student outcomes, support adaptive tutoring systems, and inform curriculum development [1,2].

Based on this research, Romero et al. [3] investigated the integration of data mining methods into Learning Management Systems (LMSs), such as Moodle, and demonstrated their potential to enhance personalized learning and instructional support. In parallel, the implementation of artificial intelligence (AI) tools, including chatbots, has increased in educational environments, with research indicating enhancements in student participation and academic achievement [4].

A challenge in this area is that academic tools frequently produce large volumes of data without corresponding outcome labels. Data annotation is time-consuming and expensive because it requires manual input from educators or domain specialists, as noted in [5]. This constraint limits the applicability of traditional supervised learning methods and motivates the use of alternative approaches, such as semi-supervised learning (SSL), which can leverage both labeled and unlabeled data to improve model performance. SSL allows models to extract underlying patterns from unlabeled data, which reduces the need for extensive manual annotation [6]. Moreover, this approach is particularly well suited to educational environments where labeled examples are frequently in short supply [7].

In SSL-based scenarios, feature selection (FS) plays a critical role in identifying the most informative variables and excluding variables that add limited value to the prediction process. FS can improve model interpretability, reduce its complexity, and enhance classification performance by shrinking the dimension of the feature space [8,9]. However, this process may also omit variables that, although weakly informative in isolation, could contribute to meaningful interactions when combined with others. Therefore, important patterns may be overlooked, potentially limiting the model's capacity to capture complex relationships in educational data. This trade-off motivates the exploration of approaches that can learn complex dependencies directly from raw data, such as Deep Learning (DL).

Recent advances in DL have demonstrated the potential to learn hierarchical representations automatically from raw data [10]. Deep neural networks (DNNs), a category of artificial neural networks (ANNs), are particularly effective in modeling complex and nonlinear relationships, making them suitable for analyzing detailed educational datasets [11]. Moreover, unlike traditional FS methods, ANN architectures can autonomously identify the most relevant features through backpropagation and optimization [12]. Nevertheless, designing and optimizing DL models remains a challenging task that requires expertise in hyperparameter tuning, architecture selection, and feature engineering.

In response to these challenges, the development of Automated Machine Learning (AutoML) aims to simplify the process of building ML models by automating tasks. Traditionally, ML and DL workflows require extensive expertise and manual efforts, making them less accessible to researchers with limited modeling engineering backgrounds. To overcome this limitation, AutoML frameworks explore a range of algorithms and configurations, often employing techniques such as genetic algorithms and neural architecture search [13,14]. Specifically, in the context of EDM, AutoML can enhance predictive modeling by optimizing feature representations and adapting to the complexity of learning behaviors, thereby improving decision-making processes in adaptive learning environments [15,16].

To evaluate these techniques in realistic educational scenarios, researchers often rely on structured datasets that contain detailed records of student interactions. One such dataset is Experience API (xAPI), which structures learning records according to a standardized framework [17]. Unlike traditional LMS logs, xAPI enables the collection of learning records across a wide variety of settings, including simulations, mobile devices, and offline activities, thus providing a more comprehensive insight into student behavior [18]. Due to its thorough coverage and depth of behavioral information, the xAPI dataset is a suitable resource for applying and assessing ML techniques in educational data analysis contexts.

Given this context, the main goal of this study is to explore the effectiveness of a fully automated DL framework pipeline for student performance prediction, leveraging AutoML to optimize model selection, hyperparameter tuning, and feature processing. We compare this approach with a recent semi-supervised FS method proposed by Yu et al. [19], using the same dataset and evaluation conditions. While the semi-supervised method enhances interpretability, our findings indicate that the AutoML-optimized DL framework

demonstrated better predictive performance under the tested conditions, likely due to its ability to uncover nonlinear patterns and complex interactions within the data [10,17,20].

To promote transparency and reproducibility, all implementation details and results of this study are publicly available at GitHub (version 1) (https://github.com/jalagrod/DL-ensemble-models-in-EDM (accessed on 15 April 2025)).

The remainder of this paper is structured as follows: Section 2 provides an overview of related work and establishes the context and relevance of our study. Section 3 details the proposed methodology, including aspects such as data preprocessing, model architectures, and evaluation metrics. Subsequently, Section 4 presents and discusses the experimental results. In Section 5, the data acquisition process is described in detail. Finally, Section 6 summarizes the main conclusions and outlines potential directions for future research.

2. Related Work

In recent years, the application of advanced data mining and ML techniques in educational contexts has enabled a deeper understanding of student behavior and academic outcomes. A notable contribution in this area was made by Amrieh et al. [17], who introduced the xAPI dataset, which contains both behavioral and demographic data collected from virtual learning environments. Their study demonstrated that appropriate preprocessing techniques can significantly enhance model performance in predicting student success.

Building on these early contributions, subsequent research incorporated more sophisticated analytical frameworks. For example, Nouira et al. [21] proposed an ontology-based framework to analyze xAPI records, thereby facilitating the structured assessment of student activity and academic performance. Hazim et al. [22] highlighted the significance of preprocessing methods, including normalization and handling missing data, to improve classification accuracy. Al Fanah and Ansari [23] examined the application of ensemble learning to xAPI data by combining multiple classifiers, which improved prediction robustness in e-learning behavior analysis.

Complementing these modeling approaches, Mihaescu and Popescu [24] surveyed public EDM datasets and underscored the flexibility of xAPI in supporting various educational analytics tasks, from dropout prediction to performance assessment. Similarly, Panda and AlQaheri [25] integrated process mining techniques to derive actionable insights from learning behavior, and Mohammad et al. [26] evaluated several ML models and highlighted the effectiveness of ensemble-based approaches such as random forests and gradient boosting.

More recently, Yu et al. [19] proposed a semi-supervised FS framework based on generalized linear regression to analyze the xAPI dataset. Their method constructs a selection matrix based on both labeled and unlabeled instances, assigns feature importance scores, and produces a ranked list of relevant attributes. The proposed method requires fewer labeled samples while maintaining competitive accuracy and outperforms baseline FS and SSL techniques. Their research confirmed the central role of behavioral features such as content interaction and participation frequency in predicting academic achievement.

In contrast to previous studies that focused on isolated components, such as data preparation, ontology modeling, or traditional ensemble methods, this study introduces an end-to-end DL framework optimized via AutoML. This proposal retains all available input features, allowing DL models to autonomously identify the most relevant patterns and interactions in the data. The methodology is described in the following section.

3. Proposed Methodology

This section presents the original methodological framework developed to predict student academic performance using an ensemble-based DL approach optimized through

AutoML. The remainder of this section details the AutoML framework used, the preprocessing strategy, model configurations, and evaluation metrics.

3.1. Research Goal and Motivation

The objective of this study is to design a reproducible and fully automated DL pipeline for student performance prediction, where model selection, hyperparameter tuning, and feature engineering are optimized through AutoML. The proposed framework is intended to reduce the need for manual feature engineering and model tuning, thereby enabling a fully automated process that adapts to the structure of educational data.

The structure of the full pipeline is illustrated in Figure 1, summarizing the automated stages involved.

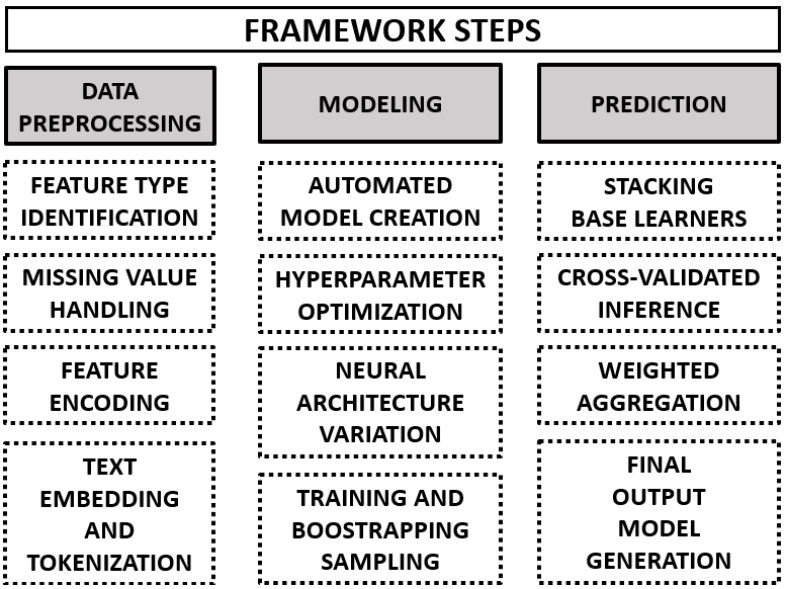

Figure 1. An overview of the proposed AutoML-enhanced DL framework for student performance prediction. The pipeline is structured into three main stages: (i) data preprocessing, which includes feature typing, encoding, and handling of missing values; (ii) modeling, which automates model creation, hyperparameter tuning, and architecture variation via bootstrapped training; and (iii) prediction, which performs ensemble stacking, cross-validated inference, and final output generation through weighted aggregation.

3.2. Automated Deep Learning with AutoGluon

AutoGluon [27] is an open-source AutoML toolkit designed to automate the development, tuning, and ensemble of machine learning and Deep Learning models. Unlike other AutoML frameworks that focus primarily on tree-based methods, AutoGluon offers native support for DNN, making it particularly suitable for this study [28]. It automates critical tasks, such as data preprocessing, model selection, hyperparameter tuning, and ensemble construction.

AutoGluon was selected over other frameworks, such as TPOT [29–31] and H2O.ai [32], due to its strong performance on tabular data and integrated support for DNN architectures.

Internally, AutoGluon employs advanced techniques, such as multilayer stacking, repeated k-fold bagging, and adaptive hyperparameter optimization. These capabilities are particularly useful in educational domains, where datasets often exhibit sparsity or class imbalance, and they contribute to building highly adaptive and generalizable models.

3.2.1. Data Preprocessing

The preprocessing step is fully automated, minimizing the need for manual intervention and ensuring consistency across training folds. AutoGluon first determines the predictive task and categorizes each input feature into one of four types: numerical, categorical, text, or temporal.

For numeric features, standardization or normalization is applied as appropriate. Categorical variables are encoded using one-hot or ordinal encoding, and missing values are handled using statistical imputation or assigned to an "unknown" category. Textual features are processed using natural language processing (NLP) techniques, such as *n*-gram tokenization and embedding generation, while temporal variables are converted into numerical format. Features deemed uninformative (e.g., unique identifiers) are automatically excluded from the pipeline.

3.2.2. Modeling and Ensemble of Deep Neural Networks

After preprocessing, AutoGluon trains a collection of base DNN models using a multilayer stacking strategy. To focus exclusively on DL, model selection is restricted to NeuralNet-Torch and NeuralNet-FastAI. This ensures that all ensemble components are based on neural architectures, excluding tree-based and linear alternatives.

The modeling process involves several stages of automation, illustrated in Figure 1, including (i) automated model creation, (ii) hyperparameter tuning via Bayesian optimization [33], random search [34], and adaptive early stopping, (iii) architectural variation using different depths, dropout rates, and activation functions [35,36], and (iv) training and bootstrapping sampling.

As shown in Figure 2, each base learner is trained on a bootstrap sample to promote diversity and reduce overfitting [37], increasing ensemble robustness through variance [38,39]. AutoGluon constructs an ensemble of up to 25 models, limited to five instances per backend. Their predictions are concatenated and passed to higher-level stacker networks, which may themselves be DNNs. These stackers, which can be homogeneous or heterogeneous classifiers [40,41], apply soft or weighted voting [42] to produce the final prediction, capturing diverse data patterns and enhancing generalization. The stacking architecture can also be extended to multiple layers to form deep ensembles in which outputs from one level are fed into the next [35]. To further improve robustness, repeated *k*-fold bagging is applied internally, training base learners on resampled subsets and providing unbiased inputs for stackers.

3.2.3. Prediction Step

To rigorously evaluate the generalization capability of the DL ensemble, the framework incorporates a two-level resampling strategy. Internally, AutoGluon applies bagging during training, which introduces variance and promotes robust learning through repeated sampling of the training data. Externally, a stratified 10-fold cross-validation (CV) procedure is employed prior to prediction. In each iteration, the dataset is partitioned into ten folds while preserving the class distribution. Nine folds are used for model training, and the remaining fold serves as a test set to ensure no data leakage.

This combination of internal bagging and external stratified CV provides a robust estimate of model performance across diverse data partitions and helps mitigate the risk of overfitting.

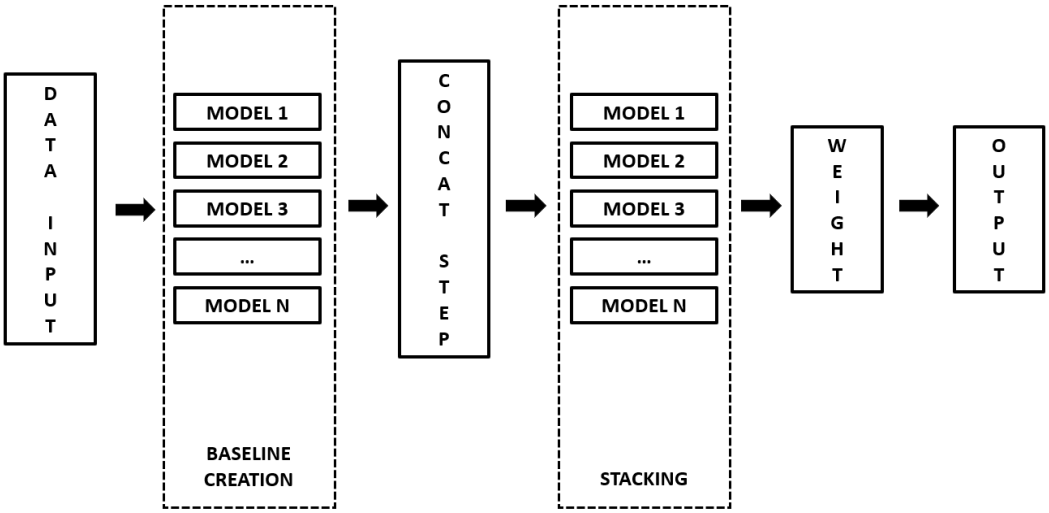

Figure 2. The ensemble process for DNNs. The ensemble integrates multiple DNN models trained on diverse subsets or configurations into a cohesive framework via stacking. Predictions are aggregated through weighted or soft voting to improve robustness and generalization.

3.2.4. Evaluation Metrics

To evaluate the predictive performance of the ensemble models, we employed several widely recognized classification metrics, including accuracy [43], precision [44], recall [45], and the F1-score [46]. Collectively, these metrics provide a comprehensive assessment of the model's classification capability, reflecting both the correctness (precision) and completeness (recall) of the predictions.

In addition to these metrics, we calculated the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) [47] to measure the discriminative ability of the proposed models. The ROC curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification thresholds, providing insight into the model's capacity to distinguish between positive and negative classes [48]. The AUC numerically summarizes this performance, taking values from 0.5 (equivalent to random classification) to 1.0 (indicating perfect class discrimination). Unlike accuracy, the ROC-AUC metric is independent of the selected classification threshold and provides robustness against imbalanced class distributions; thus, it is particularly suitable for evaluating and comparing classification models in contexts where classes may not be equally represented [47].

3.3. Experimental Procedure

To evaluate the practical benefits of the proposed framework, we compare it against the semi-supervised FS method introduced by Yu et al. [19], using the same xAPI dataset and identical experimental conditions. Their method leverages both labeled and unlabeled data to rank feature relevance and improve classification accuracy. By adopting identical evaluation settings, we ensure a fair comparison with our proposed AutoML-enhanced DL framework. Although semi-supervised FS methods offer interpretability through dimensionality reduction, they may overlook complex feature interactions; however, DL models can autonomously learn hierarchical representations from raw data. This comparison highlights the trade-offs between automation, interpretability, and predictive performance in real-world EDM scenarios.

Further details about the dataset, its structure, and feature distribution are presented in Section 4.

4. Data Collection

In this comparison, we use the xAPI dataset originally introduced by Amrieh et al. [17]. It comprises records from 480 students collected via an LMS and structured according to the xAPI standard, allowing detailed tracking of student interactions and behaviors.

The dataset contains 16 input features that are grouped into three clearly defined categories:

- Demographic features: attributes such as `gender`, `NationaliTy`, `PlaceofBirth`, `StageID`, `Relation`, and `ParentAnsweringSurvey` describe students' personal backgrounds and social contexts.
- Academic features: variables like `GradeID`, `SectionID`, `Topic`, and `Semester` represent academic contexts and settings in which student learning occurs.
- Behavioral features: engagement indicators, including `Raisedhands`, `VisitedResources`, `AnnouncementsView`, and `Discussion`, quantify the extent and type of student interactions within the LMS.

The distribution and relevance of these features are illustrated in Figures 3 and 4. These visualizations highlight important patterns in the data, such as the variability in student engagement and demographic distributions, providing an initial indication of their potential predictive power regarding academic performance.

The target variable, `Class`, categorizes student performance into three distinct groups: `High`, `Middle`, and `Low`. The distribution of the `Class` variable reveals an imbalance among the performance categories, with a predominance of students in the `Middle` class, followed by `High` and `Low`. This imbalance underscores the necessity of employing robust evaluation metrics beyond simple accuracy.

The preprocessing and encoding of features were conducted automatically via AutoGluon's built-in pipeline; however, no missing values were found in the dataset, thereby eliminating the need for specific imputation methods. The behavioral features, especially `Raisedhands`, `VisitedResources`, and `Discussion`, demonstrate strong correlations with academic outcomes, reinforcing their significance as predictors within the model (see Figure 4).

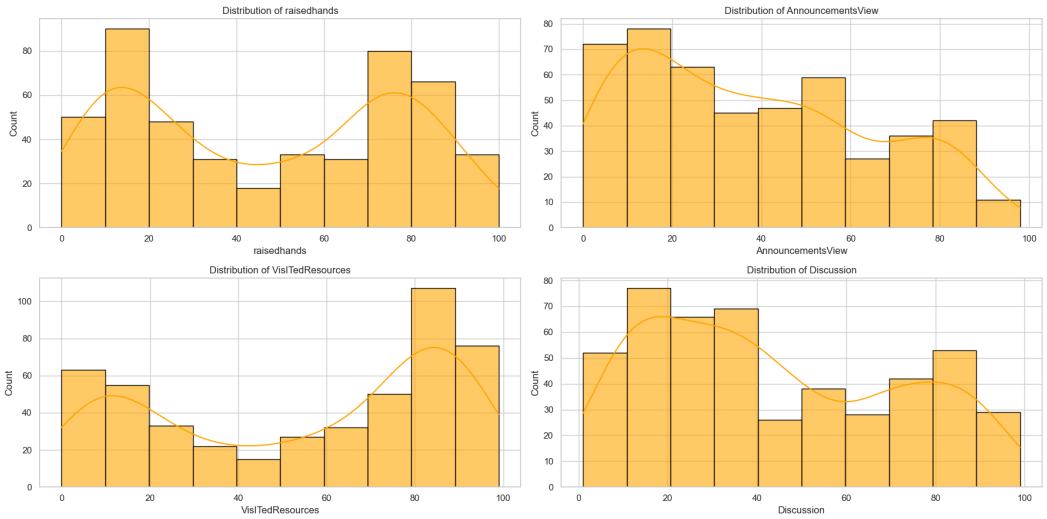

Figure 3. Distribution of categorical attributes in the dataset, including demographic (e.g., `gender`, `Nationality`), parental (e.g., `Relation`, `ParentschoolSatisfaction`), and academic context variables (e.g., `StageID`, `Semester`, `StudentAbsenceDays`). The frequency variations provide insight into the population's heterogeneity.

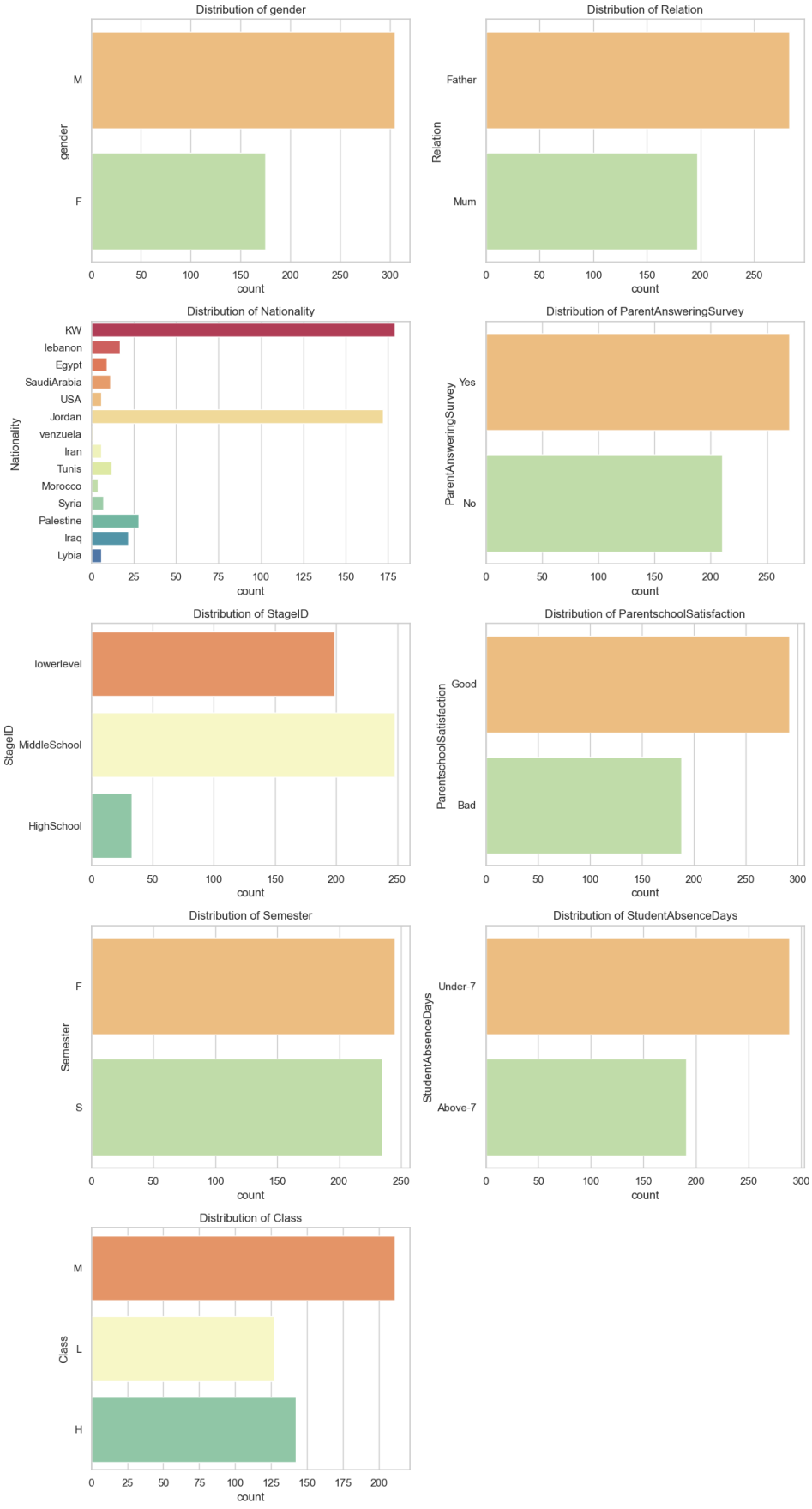

Figure 4. Distribution of key student interaction metrics, including `Raisedhands`, `AnnouncementsView`, `VisitedResources`, `Discussion`, and `Class`. These features exhibit multimodal patterns, asymmetries, and the presence of distinct behavioral profiles.

5. Results

This section presents a comparative analysis of the performance metrics obtained by the proposed DL-based approach and those reported by Yu et al. [19]. The main classification metrics are summarized in Table 1.

Table 1. A comparison of classification results on the *xAPI* dataset between Yu et al. [19] and our DL proposal. The ROC-AUC was not reported in Yu et al.'s study.

Metric	Yu et al. [19]	DL Proposal
Accuracy	0.7646	0.9548
F-score	0.6216	0.9522
Precision	0.6165	0.9597
Recall	0.6277	0.9492
ROC-AUC	Not Reported	0.9523

5.1. Cross-Validation Results

The detailed evaluation results are presented in Table 2.

Table 2. The performance metrics for each independent fold in the external 10-fold cross-validation. The metrics reported are accuracy, precision, recall, F1-score, and ROC-AUC.

Fold	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1	0.9433	0.9447	0.9433	0.9726	0.9815
2	0.9958	0.9448	0.9658	0.8959	0.9581
3	0.9708	0.9560	0.9008	0.8959	0.9581
4	0.9400	0.9483	0.9483	0.9483	0.9620
5	0.9400	0.9600	0.9900	0.9870	0.9696
6	0.9400	0.9600	0.9900	0.9870	0.9296
7	0.9400	0.9479	0.9000	0.9466	0.9167
8	0.9500	0.9989	0.9600	0.9922	0.9802
9	0.9692	0.9670	0.9692	0.9666	0.9625
10	0.9733	0.9333	0.9333	0.9333	0.9279
Average	0.9548	0.9597	0.9492	0.9522	0.9523

In these experiments, the DNN ensemble achieved a final ROC-AUC of 0.9523. The high ROC-AUC demonstrates that the model provides robust probability estimates, effectively differentiating students at higher risk of poor performance from those at lower risk. By leveraging a diverse set of architectures, the ensemble-based DL approach enhances generalizability and minimizes errors across different class distributions. The high ROC-AUC further supports the reliability and robustness of the proposed approach [49].

5.2. Statistical Significance of Performance Differences

To assess whether the performance gains achieved by our approach over the method proposed by Yu et al. [19] are statistically significant, we conducted hypothesis testing across multiple evaluation metrics.

Yu et al. [19] reported only aggregate scores without per-fold results, which precludes a direct two-sample comparison. As an alternative, we applied one-sample t-tests [50] using the per-fold metrics obtained in our experiments. These values were compared against the single-point values published in their study for the shared metrics: accuracy, precision, recall, and F1-score.

Since Yu et al. [19] did not report ROC-AUC values, this metric was excluded from the statistical analysis to ensure fair comparison.

The results indicate statistically significant improvements in all evaluated metrics ($p < 0.001$), with 95% confidence intervals from the proposed method consistently exceeding the reported values. These findings suggest that the performance gains observed in the proposed AutoML-enhanced DL framework are unlikely to be due to chance.

6. Conclusions and Future Work

This study introduces a fully automated DL framework optimized using AutoML to predict student academic performance in virtual learning environments. The primary objective was to develop a scalable and reproducible pipeline capable of autonomously learning complex patterns from raw educational data without requiring manual feature engineering or architecture tuning. To assess the practical effectiveness of the proposed method, we compared it to the semi-supervised FS framework proposed by Yu et al. [19] under identical evaluation conditions using the xAPI dataset.

The results demonstrate that the proposed DL framework consistently outperforms the FS approach across multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These improvements are statistically significant and suggest that ensemble-based DL models, when combined with AutoML techniques, offer compelling alternatives to traditional FS strategies, particularly in educational contexts where labeled data are scarce and feature interactions are complex.

A key strength of the proposed framework lies in its ability to autonomously extract hierarchical representations from behavioral indicators, such as resource usage frequency and participation patterns. This led to more robust and nuanced classification of students into performance categories. Additionally, the use of stratified 10-fold CV multilayer stacking and adaptive early stopping mechanisms helped ensure the model's generalizability and reduced the risk of overfitting.

6.1. Limitations and Future Directions

Despite its promising performance, the proposed framework has several limitations. First, the evaluation was limited to the xAPI dataset, which may restrict the generalizability of the results to other learning environments or academic domains. Second, although AutoML simplifies the modeling pipeline, the resulting DL ensembles are computationally demanding, which may hinder real-time deployment in resource-constrained settings. Third, while our approach prioritizes predictive performance, it still lacks transparency in decision-making, a critical factor in educational applications.

Therefore, future research should focus on several directions. First, we aim to extend the proposed framework to other publicly available EDM datasets and domains to assess its broad applicability. Second, we plan to integrate explainable AI (XAI) techniques into the pipeline to improve transparency and support educational decision-making. Third, we explore more efficient modeling strategies to improve interpretability and reduce training overhead.

6.2. Sustainability Considerations

Finally, the environmental impact of large-scale DL models should not be overlooked. The training and inference phases of ensemble-based DL frameworks significantly contribute to energy consumption and carbon emissions [51,52]. Future work will explore strategies for energy-efficient learning, including quantization, pruning, and the use of green infrastructure powered by renewable energy [53,54].

By addressing these limitations, this research aims to contribute meaningfully to the development of automated, interpretable, and scalable DL models for EDM that can support personalized learning and informed educational interventions.

Author Contributions: Conceptualization, J.A.L.R. and N.D.-D.; methodology, J.A.L.R. and C.D.B.G.; validation, J.A.L.R. and N.D.-D.; formal analysis, J.A.L.R. and C.D.B.G.; investigation, J.A.L.R.; writing—original draft preparation, J.A.L.R.; writing—review and editing, N.D.-D. and C.D.B.G.; visualization, J.A.L.R.; supervision, N.D.-D. and C.D.B.G.; project administration, N.D.-D. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for open access publishing: PAIDI: Intelligent Data Analysis (TIC 239) and Cátedra de Educación en Tecnologías Emergentes, Gamificación e Inteligencia Artificial (EduEmer).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in GitHub at https://github.com/jalagrod/DL-ensemble-models-in-EDM (accessed on 15 April 2025).

Acknowledgments: The authors gratefully acknowledge the financial support provided for open access publishing by the PAIDI Research Group on Intelligent Data Analysis (TIC 239) and the Education in Emerging Technologies, Gamification, and Artificial Intelligence (EduEmer).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	artificial intelligence
ANN	artificial neural network
AutoML	Automated Machine Learning
CV	cross-validation
DL	Deep Learning
DP	D'Agostino–Pearson (Test)
DNN	deep neural network
EDM	Educational Data Mining
FPR	False Positive Rate
LMS	Learning Management System
ML	machine learning
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
SSL	semi-supervised learning
TPR	True Positive Rate
XAI	Explainable Artificial Intelligence
xAPI	Experience API

References

1. Romero, C.; Ventura, S. Educational Data Mining: A Review of the State of the Art. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2010**, *40*, 601–618. [CrossRef]
2. Bakhshinategh, B.; Zaiane, O.R.; Elatia, S.; Ipperciel, D. Educational data mining applications and tasks: A survey of the last 10 years. *Educ. Inf. Technol.* **2018**, *23*, 537–553. [CrossRef]
3. Romero, C.; Ventura, S. Data Mining in Education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 12–27. [CrossRef]
4. Moral-Sánchez, S.N.; Ruiz Rey, F.J.; Cebrián-de-la Serna, M. Analysis of artificial intelligence chatbots and satisfaction for learning in mathematics education. *Int. J. Educ. Res. Innov.* **2023**, *20*, 1–14.
5. Ahmed, M.R.; Tahid, S.; Mitu, N.A.; Kundu, P.; Yeasmin, S. A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In Proceedings of the International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020; pp. 1–6.
6. Zhu, X. *Semi-Supervised Learning Literature Survey*; Technical Report 1530; University of Wisconsin-Madison: Madison, WI, USA, 2009.
7. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
8. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

9. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
11. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
12. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
13. Hutter, F.; Kotthoff, L.; Vanschoren, J. *Automated Machine Learning: Methods, Systems, Challenges*; Springer: Cham, Switzerland, 2019.
14. He, X.; Zhao, K.; Chu, X. Automl: A survey of the state-of-the-art. *Knowl.-Based Syst.* **2021**, *212*, 106622. [CrossRef]
15. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.T.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2015**, *28*, 2962–2970.
16. Vanschoren, J. Meta-Learning: A Survey. *arXiv* **2018**, arXiv:1810.03548.
17. Amrieh, E.A.; Hamtini, T.; Aljarah, I. Preprocessing and analyzing educational data set using X-API for improving student's performance. In Proceedings of the 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, Amman, Jordan, 3–5 November 2015.
18. Kevan, J.M.; Ryan, P.R. Experience API: Flexible, Decentralized and Activity-Centric Data Collection. *Technol. Knowl. Learn.* **2016**, *21*, 143–149. [CrossRef]
19. Yu, S.; Cai, Y.; Pan, B.; Leung, M.F. Semi-Supervised Feature Selection of Educational Data Mining for Student Performance Analysis. *Electronics* **2024**, *13*, 659. [CrossRef]
20. Cornillez, E.E.C. Modeling the Relationship among Digital Demographic Characteristics, Digital Literacy, and Academic Performance of Mathematics Major Students in an Online Learning Environment. *Int. J. Educ. Res. Innov.* **2024**, *21*, 1–23.
21. Nouira, A.; Cheniti-Belcadhi, L.; Braham, R. An ontology-based framework of assessment analytics for massive learning. *Comput. Appl. Eng. Educ.* **2019**, *27*, 1427–1440. [CrossRef]
22. Hazim, L.R.; Abdullah, W.D. Characteristics of data mining by classification educational dataset to improve student's evaluation. *J. Eng.* **2021**, *16*, 2825–2844.
23. Al Fanah, M.; Ansari, M.A. Understanding e-learners' behavior using data mining techniques. In Proceedings of the 2019 International Conference on Big Data, Angeles, CA, USA, 9–12 December 2019.
24. Mihaescu, M.C.; Popescu, P.S. Review on publicly available datasets for educational data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, e1403. [CrossRef]
25. Panda, M.; AlQaheri, H. An education process mining framework: Unveiling meaningful information for understanding students' learning behavior and improving teaching quality. *Information* **2022**, *13*, 29. [CrossRef]
26. Mohammad, A.S.; Al-Kaltakchi, M.T.S.; Al-Ani, J.A. Comprehensive evaluations of student performance estimation via machine learning. *Mathematics* **2023**, *11*, 3153. [CrossRef]
27. Erickson, N.; Mueller, J.; Deng, H.; Guo, A.; Li, M.; Smola, A.; Zhang, Z. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv* **2020**, arXiv:2003.06505.
28. Ferreira, L.; Pilastri, A.; Martins, C.M.; Pires, P.M.; Cortez, P. A comparison of AutoML tools for machine learning, deep learning and XGBoost. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Virtual, 18–22 July 2021; pp. 1–8.
29. Olson, R.S.; Urbanowicz, R.J.; Andrews, P.C.; Lavender, N.A.; Kidd, L.C.; Moore, J.H. Chapter Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, 30 March–1 April 2016*; Proceedings, Part I; Springer International Publishing: Cham, Switzerland, 2016; pp. 123–137. [CrossRef]
30. Le, T.T.; Fu, W.; Moore, J.H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* **2020**, *36*, 250–256. [CrossRef] [PubMed]
31. Olson, R.S.; Bartley, N.; Urbanowicz, R.J.; Moore, J.H. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. In Proceedings of the Genetic and Evolutionary Computation Conference 2016, GECCO '16, Denver, CO, USA, 20–24 July 2016; pp. 485–492. [CrossRef]
32. LeDell, E.; Poirier, S. H_2O AutoML: Scalable Automatic Machine Learning. In Proceedings of the 7th ICML Workshop on Automated Machine Learning (AutoML), Vienna, Austria, 12–18 July 2020.
33. Victoria, A.H.; Maragatham, G. Automatic tuning of hyperparameters using Bayesian optimization. *Evol. Syst.* **2021**, *12*, 217–223. [CrossRef]
34. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
35. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
36. Jain, A.; Kumar, A.; Susan, S. Evaluating deep neural network ensembles by majority voting cum meta-learning scheme. In Proceedings of the 3rd International Conference on Advances in Signal Processing and Artificial Intelligence, Porto, Portugal, 17–19 November 2021; Springer: Cham, Switzerland, 2022.

37. Kim, H.; Kim, H.; Moon, H.; Ahn, H. A weight-adjusted voting algorithm for ensembles of classifiers. *J. Korean Stat. Soc.* **2011**, *40*, 437–449. [CrossRef]
38. West, D.; Dellana, S.; Qian, J. Neural network ensemble strategies for financial decision applications. *Comput. Oper. Res.* **2005**, *32*, 2543–2559. [CrossRef]
39. Zhou, Z.; Wu, J.; Tang, W. Ensembling neural networks: Many could be better than all. *Artif. Intell.* **2002**, *137*, 239–263. [CrossRef]
40. Ganaie, M.; Hu, M.; Malik, A.; Tanveer, M. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
41. Iqbal, N.; Jamil, H.; Kim, D. An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection. *J. Netw. Comput. Appl.* **2023**, *212*, 103560.
42. Dietterich, T. Ensemble methods in machine learning. In *Multiple Classifier Systems. MCS 2000*; Springer: Berlin/Heidelberg, Germany, 2000.
43. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufman: San Francisco, CA, USA, 2016.
44. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*; ACM Press/Addison-Wesley: New York, NY, USA, 1999.
45. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; pp. 1–482.
46. van Rijsbergen, C.J. *Information Retrieval*; Butterworth-Heinemann: Oxford, UK, 1979.
47. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
48. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
49. Zou, K.H.; O'Malley, A.J.; Mauri, L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* **2007**, *115*, 654–657. [CrossRef]
50. Student. The probable error of a mean. *Biometrika* **1908**, *6*, 1–25. [CrossRef]
51. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. *arXiv* **2019**, arXiv:1906.02243.
52. Patterson, D.; Gonzalez, J.; Le, Q.V.; Liang, C.; Munguia, L.M.; Rothchild, D.; Dean, J. Carbon Emissions and Large Neural Network Training. *arXiv* **2021**, arXiv:2104.10350.
53. Xu, C.; Wang, Z.; Qin, J.; Jiang, Y. Energy-Efficient Deep Learning Model Compression: A Survey. *Neurocomputing* **2021**, *461*, 173–196.
54. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63. [CrossRef]