

Causality-Inspired Representation Learning for Student Performance Prediction

Boya Liu^{1*,†}, Zijun Li^{2†} and Jie Yang¹

¹College of Marxism, Xi'an Eurasia University, Xi'an, Shaanxi, China

²College of Electronic and Information Engineering, Tongji University, Shanghai, China

Abstract

Predicting student performance is crucial for timely educational interventions, yet most existing approaches primarily focus on statistical pattern learning that may not fully reflect the underlying factors associated with academic success. This limitation can hinder their ability to support effective, evidence-based strategies. To address this challenge, we propose **Cavat**, a novel framework that integrates Causality-inspired feature selection via ensemble voting with attention-based representation learning. Cavat first constructs a consensus graph by aggregating multiple causal discovery algorithms (PC, FCI, GES) through a weighted voting mechanism, and uses it to identify features potentially relevant to student outcomes. It then employs a dual-branch neural architecture for multi-task prediction: a multi-head attention module learns interpretable representations from these selected features, while a parallel 1D CNN captures broader statistical patterns from the full feature set. The two representations are combined to leverage complementary perspectives for prediction. Extensive experiments show that Cavat outperforms strong baselines by **5.5%** on exam score prediction and **4.0%** on performance classification. In addition, the framework provides insights into feature importance through attention analysis, offering a more interpretable perspective for understanding student performance. These results demonstrate the potential of combining causality-inspired structure learning with representation learning for building more robust and informative educational prediction models.

Keywords

Educational Data Mining, Student Performance Prediction, Feature Engineering,

1. Introduction

Student performance prediction is a cornerstone of modern educational research. Accurate forecasting of academic outcomes empowers educators to identify at-risk students early, deploy targeted interventions, and ultimately reduce failure rates [1]. Beyond individual student support, these predictive models are critical for optimizing resource allocation, informing evidence-based policymaking, and advancing personalized learning at scale [2]. In an era of increasing educational digitalization, such capabilities are essential for enhancing teaching quality, streamlining administration, and fostering more equitable and efficient education systems.

The technical foundation for this work lies in Educational Data Mining (EDM), which applies advanced analytical techniques to diverse educational datasets [3]. Data from online platforms, institutional databases, and blended learning environments are mined to uncover patterns that can refine instructional strategies and policy [4]. While traditional machine learning methods such as Artificial Neural Networks and Support Vector Machines (SVM) have shown promise [4, 5, 6], they often rely on manual feature engineering, which can limit their ability to capture complex, high-dimensional interactions in student data. Deep learning offers a powerful alternative with its capacity for automatic feature representation [7, 8], yet its potential to incorporate more structured and interpretable modeling in EDM remains underexplored.

This challenge highlights two key limitations in existing EDM systems. First, a heavy reliance on correlational patterns may lead to the identification of relationships that do not generalize well across dif-

AI for Education Day, SIGKDD 2026, August 12, 2026, Jeju, Korea

*Corresponding author.

†These authors contributed equally.

✉ liuboya@eurasia.edu.cn (B. Liu); lizijun@tongji.edu.cn (Z. Li); yangjie@eurasia.edu.cn (J. Yang)



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

ferent contexts, potentially limiting their usefulness for intervention and decision-making [9]. Second, manual feature engineering may overlook latent pedagogical factors, introducing biases that can hinder effective personalization. Incorporating structured information about feature relationships, such as those derived from causal discovery methods [9], together with the representation learning capabilities of deep models [7], presents a promising direction for improving both robustness and interpretability.

To address this challenge, we propose **Cavat**, a novel framework that integrates **causality-inspired** feature selection via attention-based representation learning. Our approach begins by constructing a consensus graph through a weighted voting mechanism that aggregates multiple causal discovery algorithms (PC, FCI, GES), aiming to provide a more stable estimation of feature relationships. Based on this graph, we identify a subset of features that are potentially relevant to student outcomes.

We then design a dual-branch neural architecture to capture complementary feature perspectives. A multi-head attention mechanism learns representations from the selected features, modeling their interactions in a context-dependent manner, while a parallel 1D Convolutional Neural Network (CNN) captures broader statistical patterns from the full feature set. By fusing these representations, our multi-task learning framework simultaneously predicts both continuous exam scores (regression) and categorical performance levels (fail/pass/distinction).

Our contributions are summarized as follows:

- We propose a hybrid framework that integrates ensemble-based causal discovery with deep learning, introducing a weighted voting mechanism to obtain a more stable estimation of feature relationships.
- We design a dual-branch neural architecture that captures complementary feature perspectives, combining representations derived from causality-inspired feature selection and global statistical patterns to improve both performance and interpretability.
- We demonstrate the effectiveness of Cavat through extensive experiments. Our model consistently outperforms strong baselines by 5.5% on exam score prediction and 4.0% on student performance classification. Ablation studies further validate our design choices, and our analysis provides insights into feature importance under different predictive scenarios.

2. Related Work

Educational Data Mining (EDM) has extensively studied student performance prediction, evolving from traditional statistical models to deep learning and hybrid systems. While these methods have achieved strong predictive performance, most of them primarily focus on learning statistical patterns from data, with limited incorporation of structured feature relationships.

Early studies employed statistical and classical machine learning models to identify predictive factors. Regression analysis highlighted key academic and demographic variables [10], while classifiers such as Decision Trees [11, 1], SVM [12, 13, 14], KNN [13, 15], and Naive Bayes [12] achieved moderate to high accuracy using structured features [16, 17]. However, these methods rely on manual feature engineering and struggle with complex, non-linear patterns. Deep learning significantly improved predictive performance by learning representations from high-dimensional and sequential data. RNN-based models (e.g., LSTM, BiLSTM) capture temporal learning dynamics [18, 19, 20], while DNNs and attention mechanisms model heterogeneous features and salient signals [21, 22, 23]. GNN-based approaches further exploit relational structures in student-knowledge interactions [24, 25, 26]. Despite achieving strong performance, these models are typically black-box and lack interpretability, while also overlooking issues such as class imbalance.

Moreover, Knowledge tracing (KT) [27], a central task in learner modeling, aims to estimate a learner’s evolving latent knowledge state from a sequence of learning interactions and to predict future performance. Deep Knowledge Tracing [28] employs recurrent neural networks to capture temporal dependencies directly from student interaction histories. For example, SAKT [29] and AKT [30], use attention mechanisms to identify relevant prior interactions and incorporate contextual or psychometric information, improving both predictive performance and interpretability. Recent work explores hybrid

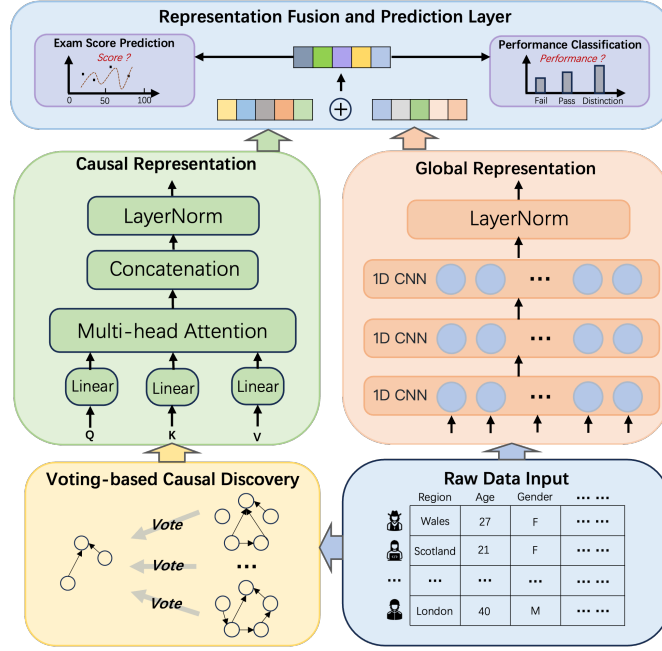


Figure 1: An overview of the Cavat framework, detailing the flow from robust causal discovery to multi-task prediction.

and ensemble methods to further boost performance. Ensemble models (e.g., Gradient Boosting, XG-Boost) improve robustness [31, 32], while hybrid approaches combine optimization techniques such as TLBO [4] and genetic algorithms [33] with traditional classifiers. More recent systems integrate deep learning, imbalance handling (e.g., SMOTEENN), and interpretability tools like SHAP [34].

While these methods have achieved strong predictive performance, most of them primarily focus on learning statistical patterns from data and Knowledge tracing primarily focus on fine-grained, sequential question-level knowledge-state estimation. In contrast, incorporating structured information about feature relationships remains relatively underexplored in EDM. This motivates our work, which aims to integrate structure learning with representation learning to improve both robustness and interpretability.

3. Method

This section details the architecture of Cavat, our proposed framework for student performance prediction. As illustrated in Figure 1, our methodology follows four key stages: (1) estimating feature relationships using a voting-based ensemble of causal discovery algorithms; (2) learning feature representations through a dual-branch neural network; (3) fusing the learned representations; and (4) performing multi-task prediction based on the fused representation.

3.1. Voting-Based Causal Discovery Framework

The goal of this component is to estimate structured relationships among features that can provide useful inductive bias for downstream learning. However, individual causal discovery algorithms such as the Peter-Clark (PC) [35] algorithm, the Fast Causal Inference (FCI) [35] algorithm and the Greedy Equivalence Search (GES) [36] algorithm are known to be sensitive to hyperparameters, dataset characteristics, and violations of their underlying assumptions. As a result, relying on a single method may lead to unstable or inconsistent structure estimation. At a high level, PC, FCI, and GES infer graphical structures among observed variables, but they differ in how they search for such structures and in the assumptions they make. PC and FCI are constraint-based methods: they begin with a densely connected graph and remove edges when conditional-independence tests indicate that two variables

become independent after conditioning on other variables. PC generally assumes causal sufficiency, meaning that no relevant latent common causes are omitted, whereas FCI relaxes this assumption and represents uncertainty caused by possible latent confounders through partially oriented edges. In contrast, GES is a score-based method that greedily adds and removes edges to optimize a model-selection score over equivalence classes of directed acyclic graphs. These methods commonly rely on assumptions such as acyclicity, the causal Markov condition, and faithfulness, and their results can be affected by finite samples and the choice of independence tests or scoring functions. Consequently, we treat their outputs as estimated structural hypotheses that provide an inductive bias for representation learning, rather than as definitive evidence of causal effects.

3.1.1. Causal Discovery Algorithm

To improve robustness, we adopt an ensemble strategy that integrates multiple causal discovery algorithms. Specifically, we combine PC, FCI, and GES, which offer complementary perspectives on structure learning, and aggregate their outputs to form a consensus graph. This approach is motivated by the idea that combining multiple estimators can help reduce method-specific biases and improve stability. The resulting graph is used as a structural prior to identify a subset of features that are potentially relevant to student outcomes. The focus of this work is not on improving individual causal discovery algorithms, but on leveraging their aggregated outputs to guide representation learning.

3.1.2. Voting Mechanism

Our voting mechanism synthesizes the outputs of the $k = 3$ base algorithms (PC, FCI, GES) to construct a consensus graph G^* . The overall procedure is summarized in Algorithm 1. It consists of three stages: **Edge Existence Voting**: For each potential edge between variables (X, Y) , we count how many algorithms include this edge. An undirected edge is added to G^* if the vote count is no smaller than a threshold τ (typically $\lceil k/2 \rceil$). **Edge Orientation Voting**: For each retained edge in G^* , we collect votes for its possible orientations. The orientation with the highest support is assigned if its vote count is at least τ_{orient} , where $\tau_{\text{orient}} \geq \tau$. **Conflict Resolution**: Since the aggregated graph may still contain cycles, we further enforce acyclicity by iteratively removing the edge with the weakest support from each detected cycle.

Through this procedure, the outputs of different causal discovery algorithms are aggregated into a single consensus graph, which is then used to identify a structured subset of features for downstream representation learning.

Algorithm 1 Voting Mechanism for Causal Discovery

Require: Graphs $\{G_1, \dots, G_k\}$, thresholds $\tau, \tau_{\text{orient}}$.

Ensure: Consensus graph G^* .

- 1: Initialize an empty graph G^* over all variables.
 - 2: **Edge Existence Voting**: For each pair (X, Y) , add an undirected edge to G^* if $\sum_{i=1}^k \mathbb{I}(X - Y \in G_i) \geq \tau$.
 - 3: **Edge Orientation Voting**: For each edge in G^* , assign the orientation with maximum votes if the vote count is $\geq \tau_{\text{orient}}$.
 - 4: **Conflict Resolution**: Iteratively remove the lowest-voted edge from any cycles in G^* until acyclicity is achieved.
 - 5: **return** G^*
-

3.2. Dual-Branch Feature Representation

Our representation learning framework is designed to capture complementary feature perspectives. Rather than assuming a strict separation between different types of information, we use a dual-branch architecture to model both structured feature subsets and broader statistical patterns.

3.2.1. Feature Representation from Selected Features

Let $\mathbf{X}_c \in \mathbb{R}^{n \times d_c}$ denote the subset of features selected based on the consensus graph. This branch aims to model interactions among these features. We adopt a multi-head self-attention mechanism to capture context-dependent relationships.

The attention mechanism enables the model to capture the relative importance of different features by modeling their interactions within each student's profile.

The input $\mathbf{X}_c$ is projected into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices, and the attention output is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (1)$$

The final representation $\mathbf{H}_c$ is obtained via residual connection and layer normalization:

$$\mathbf{H}_c = \text{LayerNorm}(\mathbf{X}_c + \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \quad (2)$$

3.2.2. Global Feature Representation

In parallel, we model the full feature set $\mathbf{X} \in \mathbb{R}^{n \times d}$ to capture broader statistical dependencies that may not be fully reflected in the selected subset. We employ a 1D Convolutional Neural Network (CNN), which is effective at capturing local interactions among features when treated as sequences.

The CNN extracts hierarchical patterns, and its output is combined with the original input via a residual connection and pooling operation to produce the global representation $\mathbf{H}_g$:

$$\mathbf{H}_g = \text{LayerNorm}(\mathbf{X} + \text{Linear}(\text{Pool}(\text{ConvLayers}(\mathbf{X})))) \quad (3)$$

3.3. Representation Fusion and Multi-Task Prediction

Our method can be viewed as introducing structured feature relationships as an inductive bias into representation learning.

3.3.1. Representation Fusion

To integrate the two representations, we project $\mathbf{H}_c$ and $\mathbf{H}_g$ into a shared space and combine them using a weighted sum:

$$\mathbf{H} = g \cdot \mathbf{H}'_c + (1 - g) \cdot \mathbf{H}'_g \quad (4)$$

We set $g = 0.5$ to ensure a balanced contribution from both branches. This design does not assume the dominance of either representation, but instead encourages the model to leverage complementary information from both structured feature subsets and global statistical patterns.

3.3.2. Prediction Layer

We adopt a multi-task learning framework to improve representation quality by jointly optimizing regression and classification objectives.

Exam Score Prediction (Regression):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (5)$$

Performance Classification:

$$\mathcal{L}_{\text{class}} = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^3 z_{i,k} \log(p_{i,k}) \quad (6)$$

Joint Optimization:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{class}} \quad (7)$$

where λ balances the two tasks (set to 0.1 in our experiments).

4. Experiment

4.1. Dataset

We use the Open University dataset [37]¹ The dataset is split into 80% training and 20% test sets.

4.2. Baseline and Evaluation Metrics

We compare our proposed model against the following baseline methods, which are widely used in machine learning tasks: **Support Vector Machine (SVM)** [38], **Multi-Layer Perceptron (MLP)** [39], **Decision Tree (DT)** [40], **Random Forest (RF)** [41], **XGBoost** [42]. Each baseline is trained on the raw feature set $\mathbf{X} \in \mathbb{R}^{n \times d}$, without the causal feature selection and global feature transformation applied in our model. We evaluate the model’s performance on two tasks: regression (exam score prediction) and a student performance classification task. For the exam score regression task, we predict continuous exam scores using the following metrics: **Mean Squared Error (RMSE)**, **Mean Absolute Error (MAE)** and **R-squared (R^2)**. For the student performance classification task, we use the following metrics: **Accuracy**, **Precision**, **Recall**, **F1 Score**.

4.3. Results of Exam Score Prediction Task

Table 1 presents the results for the exam score prediction task. Our proposed model outperforms all baselines across all metrics, achieving the lowest RMSE and MAE and the highest R^2 , achieving 5.5% performance improvement compared to the best baselines. This indicates that the combination of causal and global feature representations captures patterns relevant to exam score prediction.

Table 1

Regression Task Results for Exam Score Prediction

Method	RMSE ↓	MAE ↓	R^2 ↑
SVM	18.5	13.2	0.60
MLP	17.8	12.9	0.62
Decision Tree	19.5	15.4	0.54
Random Forest	18.1	13.2	0.58
XGBoost	17.5	13.0	0.62
Cavat	17.0	12.5	0.68

4.4. Results of Student Performance Classification Task

Table 2 presents the results for the student performance classification task. The proposed model achieves the best performance across all evaluation metrics, including accuracy, precision, recall, and F1-score, with a 4.0% improvement over the strongest baseline. These results indicate that the proposed framework is effective in distinguishing between fail, pass, and distinction outcomes. While the MLP and RF baselines also achieve competitive performance, the results suggest that incorporating structured feature representations can further enhance predictive performance.

4.5. Ablation Study

To assess the contribution of each component in our proposed model, we conduct a comprehensive ablation study by removing or modifying key modules: (1) the causal feature representation ($\mathbf{H}_c$), (2) the global feature representation ($\mathbf{H}_g$), and (3) the causal discovery strategy. Specifically, **Cavat-NoCausal** removes the causal branch and relies solely on the global representation $\mathbf{H}_g$, while **Cavat-NoGlobal** removes the global branch and uses only the representation learned from features identified

¹Our used dataset is available at <https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad>, note that we select a subset of the whole dataset in our paper.

Table 2

Classification Task Results for Student performance

Method	Accuracy $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F1-score $\uparrow$
SVM	0.69	0.68	0.69	0.68
MLP	0.71	0.70	0.71	0.71
Decision Tree	0.67	0.68	0.67	0.68
Random Forest	0.76	0.75	0.76	0.76
XGBoost	0.83	0.82	0.80	0.81
Cavat	0.87	0.85	0.83	0.84

by causal discovery ($\mathbf{H}_c$). In addition, we evaluate three variants using individual causal discovery algorithms without the proposed voting mechanism, namely **Cavat-PC**, **Cavat-FCI**, and **Cavat-GES**. These variants allow us to examine the impact of different causal structure estimation strategies. Table 3 reports the results on both regression and classification tasks. We observe that the full model consistently outperforms **Cavat-NoCausal**, indicating that incorporating features identified by causal discovery provides useful signals for prediction. Meanwhile, the performance gap between **Cavat** and **Cavat-NoGlobal** suggests that modeling broader statistical dependencies from the full feature space also plays a crucial role. These results support our design choice of combining two complementary representation perspectives. Furthermore, comparing **Cavat** with **Cavat-PC**, **Cavat-FCI**, and **Cavat-GES**, we find that the proposed voting-based causal discovery framework consistently yields better performance than any single algorithm. Notably, these improvements are consistent across both regression and classification tasks, suggesting that the voting mechanism provides a more robust and stable estimation of informative feature structures under different prediction objectives. Overall, the ablation study demonstrates that both representation branches contribute complementary inductive biases, and that the voting-based causal discovery mechanism further enhances model robustness.

Table 3

Ablation Study Results

Method	Exam Score Prediction			Student Performance Classification			
	RMSE $\downarrow$	MAE $\downarrow$	R^2 $\uparrow$	Accuracy $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F1-score $\uparrow$
Cavat-NoCausal	18.7	14.0	0.61	0.82	0.80	0.78	0.79
Cavat-NoGlobal	20.9	16.3	0.52	0.80	0.79	0.77	0.78
Cavat-PC	18.6	13.8	0.62	0.83	0.81	0.79	0.80
Cavat-FCI	18.3	13.3	0.64	0.84	0.82	0.81	0.81
Cavat-GES	18.4	13.3	0.63	0.83	0.82	0.81	0.81
Cavat	17.0	12.5	0.68	0.87	0.85	0.83	0.84

4.6. Analysis of Causal Feature Attention Weights

To enhance interpretability, we analyze the attention weights of causal features identified by the voting-based causal discovery framework (i.e., Region, IMD, Assignment4–6 scores, and Engagement level) under different prediction settings. The attention weights reflect the relative importance of each causal feature. By examining these weights, we analyze feature contributions under different prediction settings (e.g., pass vs. distinction). We extract attention scores that are computed via $\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$, where $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{n \times d_k}$ are derived from $\mathbf{X}_c$. These scores capture pairwise feature contributions. To obtain class-level feature importance (e.g., “pass” vs. “distinction”), we average attention scores across all heads and samples within each class, and compute the mean attention weight for each causal feature. The resulting normalized vector reflects the relative importance of each feature.

We visualize the learned attention weights of the selected features under different prediction settings using bar plots, as shown in Figure 2. The plots present normalized attention weights, computed

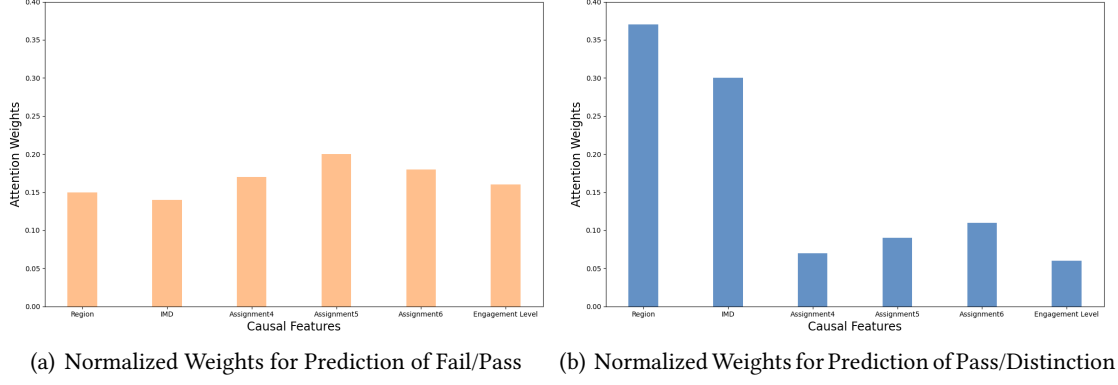


Figure 2: The Learnt Attention Weights of Causal Features for Different Prediction Settings

separately for each class to facilitate comparison. The visualization reveals different patterns of feature importance across tasks. As shown in Figure 2(a), the attention weights for the “Fail vs. Pass” setting (where “Distinction” is grouped with “Pass”) are relatively evenly distributed across features. This suggests that the model leverages multiple features jointly when distinguishing between failing and passing students. Among them, Assignment5 and Assignment6 receive slightly higher attention, indicating that these features may play a relatively stronger role in the model’s decision process. In contrast, background-related variables such as Region and IMD receive comparatively lower attention in this setting.

In comparison, Figure 2(b) shows a more uneven distribution of attention weights for the “Pass vs. Distinction” setting. In this case, Region and IMD receive higher attention weights, while assignment scores (Assignment4, Assignment5, Assignment6) and Engagement Level receive lower attention. This suggests that the model relies more on background-related features when distinguishing higher-performing students, whereas academic features contribute less prominently in this scenario.

Overall, the attention patterns differ across prediction tasks, indicating that the model adapts its focus depending on the classification objective. For the pass–fail setting, academic-related features appear to contribute more strongly, while for the pass–distinction setting, background-related features receive relatively higher emphasis. These observations provide an interpretable view of how different feature groups are utilized by the model. It is important to note that attention weights reflect the model’s internal feature weighting rather than causal effects. Nevertheless, this analysis offers useful insights into the model’s behavior and highlights how incorporating structured feature selection can support more interpretable predictions in educational settings.

5. Conclusion

In this paper, we study the problem of improving prediction robustness and interpretability in Educational Data Mining. We proposed **Cavat**, a novel framework that integrates **Causality**-inspired feature selection via ensemble voting with **attention**-based representation learning. Our Cavat constructs a consensus graph that provides a more stable estimation of feature relationships. The dual-branch architecture captures both consensus structure feature and statistical patterns from the full feature set perspectives. Experimental results show Cavat consistently outperforms strong baselines and the ablation studies further validate our design choices. In addition to predictive performance, our framework provides insights into feature importance through attention-based analysis. The results suggest that different features (e.g., academic-related and socioeconomic variables) contribute differently under various prediction settings, offering a more novel perspective for understanding student performance.

For future work, we plan to further investigate more principled strategies for combining and weighting different structure learning methods, as well as exploring more rigorous ways to incorporate causal assumptions into representation learning.

Declaration on Generative AI

During the preparation of this work, the authors used DeepSeek in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] M. Pandey, V. K. Sharma, A decision tree algorithm pertaining to the student performance analysis and prediction, *International Journal of Computer Applications* 61 (2013) 1–5.
- [2] R. S. Baker, et al., Challenges for the future of educational data mining: The baker learning analytics prizes, *Journal of educational data mining* 11 (2019) 1–17.
- [3] Y. Lin, H. Chen, W. Xia, F. Lin, Z. Wang, Y. Liu, A comprehensive survey on deep learning techniques in educational data mining, *arXiv preprint arXiv:2309.04761* (2023).
- [4] M. Arashpour, E. M. Golafshani, R. Parthiban, J. Lamborn, A. Kashani, H. Li, P. Farzanehfard, Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization, *Computer Applications in Engineering Education* 31 (2023) 83–99.
- [5] W. Chen, Z. Shen, Y. Pan, K. Tan, C. Wang, Applying machine learning algorithm to optimize personalized education recommendation system, *Journal of Theory and Practice of Engineering Science* 4 (2024) 101–108.
- [6] G. Li, J. Shuai, Y. Hu, Y. Zhang, Y. Wang, T. Yang, N. Xiong, Dkt-lcirt: A deep knowledge tracking model integrating learning capability and item response theory, *Electronics* 11 (2022) 3364.
- [7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *nature* 521 (2015) 436–444.
- [8] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, B. Navarro-Colorado, A systematic review of deep learning approaches to educational data mining, *Complexity* 2019 (2019) 1306039.
- [9] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2009.
- [10] A. Goold, R. Rimmer, Indicators of performance in first-year computing, in: *Proceedings 23rd Australasian Computer Science Conference. ACSC 2000* (Cat. No. PR00518), IEEE, 2000, pp. 74–80.
- [11] R. Asif, A. Mercer, M. K. Pathan, Investigating performance of students: a longitudinal study, in: *Proceedings of the fifth international conference on learning analytics and knowledge*, 2015, pp. 108–112.
- [12] M. Kumar, A. Singh, Evaluation of data mining techniques for predicting student's performance, *International Journal of Modern Education and Computer Science* 8 (2017) 25–31.
- [13] S. Wiyono, D. S. Wibowo, M. F. Hidayatullah, D. Dairoh, Comparative study of knn, svm and decision tree algorithm for student's performance prediction, *(IJCSAM) International Journal of Computing Science and Applied Mathematics* 6 (2020) 50–53.
- [14] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, K. Phasinam, Classification and prediction of student performance data using various machine learning algorithms, *Materials today: proceedings* 80 (2023) 3782–3785.
- [15] H. Turabieh, S. A. Azwari, M. Rokaya, W. Alosaimi, A. Alharbi, W. Alhakami, M. Alnfai, Enhanced harris hawks optimization as a feature selection for the prediction of student performance, *Computing* 103 (2021) 1417–1438.
- [16] R. J. Oskouei, M. Askari, Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies), *Computer Engineering and Applications Journal* 3 (2014) 79–88.
- [17] A. Almasri, E. Celebi, R. S. Alkhawaldeh, Emt: Ensemble meta-based tree model for predicting student performance, *Scientific Programming* 2019 (2019) 3610248.
- [18] F. Okubo, T. Yamashita, A. Shimada, H. Ogata, A neural network approach for students' performance prediction, in: *Proceedings of the seventh international learning analytics & knowledge conference*, 2017, pp. 598–599.

- [19] H. Liu, X. Chen, F. Zhao, Learning behavior feature fused deep learning network model for mooc dropout prediction, *Education and Information Technologies* 29 (2024) 3257–3278.
- [20] B. K. Yousafzai, S. A. Khan, T. Rahman, I. Khan, I. Ullah, A. Ur Rehman, M. Baz, H. Hamam, O. Cheikhrouhou, Student-performulator: student academic performance using hybrid deep neural network, *Sustainability* 13 (2021) 9775.
- [21] P. P. Bendangnuksung, Students' performance prediction using deep neural network, *International Journal of Applied Engineering Research* 13 (2018) 1171–1176.
- [22] S. Chen, Y. Ding, A machine learning approach to predicting academic performance in pennsylvania's schools, *Social Sciences* 12 (2023) 118.
- [23] Y. Chen, G. Wei, J. Liu, Y. Chen, Q. Zheng, F. Tian, H. Zhu, Q. Wang, Y. Wu, A prediction model of student performance based on self-attention mechanism, *Knowledge and Information Systems* 65 (2023) 733–758.
- [24] Y. Yang, J. Shen, Y. Qu, Y. Liu, K. Wang, Y. Zhu, W. Zhang, Y. Yu, Gikt: a graph-based interaction model for knowledge tracing, in: *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2020, Springer, 2021*, pp. 299–315.
- [25] Q. Ni, T. Wei, J. Zhao, L. He, C. Zheng, Hhskt: A learner-question interactions based heterogeneous graph neural network model for knowledge tracing, *Expert Systems with Applications* 215 (2023) 119334.
- [26] P. Hai-tao, F. Ming-qu, Z. Hong-bin, Y. Bi-zhen, L. Jin-jiao, L. Chun-fang, Z. Yan-ze, S. Rui, Predicting academic performance of students in chinese-foreign cooperation in running schools with graph convolutional network, *Neural Computing and Applications* 33 (2021) 637–645.
- [27] A. T. Corbett, J. R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge, *User modeling and user-adapted interaction* 4 (1994) 253–278.
- [28] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, J. Sohl-Dickstein, Deep knowledge tracing, *Advances in neural information processing systems* 28 (2015).
- [29] S. Pandey, G. Karypis, A self-attentive model for knowledge tracing, *arXiv preprint arXiv:1907.06837* (2019).
- [30] A. Ghosh, N. Heffernan, A. S. Lan, Context-aware attentive knowledge tracing, in: *KDD, 2020*, pp. 2330–2339.
- [31] S. B. Keser, S. Aghalarova, Hela: A novel hybrid ensemble learning algorithm for predicting academic performance of students, *Education and Information Technologies* 27 (2022) 4521–4552.
- [32] A. Asselman, M. Khaldi, S. Aammou, Enhancing the prediction of student performance based on the machine learning xgboost algorithm, *Interactive Learning Environments* 31 (2023) 3360–3379.
- [33] S. S. Shreem, H. Turabieh, S. Al Azwari, F. Baothman, Enhanced binary genetic algorithm as a feature selection to predict student performance, *Soft Computing* 26 (2022) 1811–1823.
- [34] Y. Alshamaila, H. Alsawalqah, I. Aljarah, M. Habib, H. Faris, M. Alshraideh, B. A. Salih, An automatic prediction of students' performance to support the university education system: a deep learning approach, *Multimedia Tools and Applications* 83 (2024) 46369–46396.
- [35] P. Spirtes, C. N. Glymour, R. Scheines, *Causation, prediction, and search*, MIT press, 2000.
- [36] D. M. Chickering, Optimal structure identification with greedy search, *Journal of machine learning research* 3 (2002) 507–554.
- [37] J. Kuzilek, M. Hlosta, Z. Zdrahal, Open university learning analytics dataset, *Scientific data* 4 (2017) 1–8.
- [38] J. Platt, Sequential minimal optimization: A fast algorithm for training support vector machines (1998).
- [39] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain., *Psychological review* 65 (1958) 386.
- [40] L. Ying, et al., Decision tree methods: applications for classification and prediction, *Shanghai archives of psychiatry* 27 (2015) 130.
- [41] L. Breiman, Random forests, *Machine learning* 45 (2001) 5–32.
- [42] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: *KDD, 2016*, pp. 785–794.