Mining Student Habit Patterns through Contrastive Self-Supervised Representation Learning and Kohonen Topological Clustering

Md. Tanvir Hossain
dept. of CSE
BRAC University
Dhaka, Bangladesh
tanvir.eece.mist@gmail.com

***Abstract*—Collecting and processing information about students' actions is very complex with the available data being unlabeled and high-dimensional. Contrastive self-supervised representation learning, Kohonen Self-Organizing Maps (SOM) and classical clustering algorithms are used to address the issue. Latent embeddings are created using an approach inspired by SimCLR and through applying specially designed methods such as removing or changing parts of the images. The embeddings help group similar patterns using SOM, KMeans and GMM. To obtain the clustering scores, Silhouette scores, NMI and ARI are applied, where the probability of patients developing the disease represents the ground truth. Even though the contrastive loss converges properly, the cluster details and its interaction with dropout risk suggest that the task is not easily separated due to the complexity of people's behaviors. Because it is interpretable and flexible, the approach in the framework supports creating unique education plans for students. Further efforts are being made to combine modern clustering methods and data collected using multiple means.**

***Index Terms*—Student behavior analysis, Contrastive self-supervised learning, Representation learning, Kohonen Self-Organizing Maps, Unsupervised clustering, Educational data mining, Pattern discovery, Neural networks, Deep learning, Cluster evaluation metrics**

I. Introduction

With more students relying on digital education, teachers and experts now collect data about a student's study routine, choice of resources and activity. It is important to discover relevant patterns here, as this will drive development in individual education, early intervention and support in academics. Nevertheless, supervised learning models often need too many labeled samples, something difficult or very costly to get in schools. Besides, not every hand-made feature easily accounts for how a student's behavior happens at different stages of education. Therefore, this study demonstrates that expert habit data from students can be effectively trained with an unsupervised path, based on contrastive learning. Following the SimCLR approach, the method uses task-related augmentations to make the model universal and define natural traits of animal behaviors on its own. Afterward, the compact embeddings are clustered by Kohonen SOM and by using KMeans and GMM algorithms. The two approaches, self-supervised learning and multiple clustering, help discover all the patterns and confirm their correct grouping. Observations gained from evaluating the framework's success point to the many variations among the habits of students. This technique can be used to process a wide range of data and guide teachers' decisions about personalized education.

II. Literature Review

As digital education platforms grow, there is a large amount of information on student behavior that cannot be analyzed without advanced technology. Since labeled data is not always scarce in recent times, studies have been carried out on self-supervised and contrastive learning approaches. Scarlatos et al. (2022) developed a framework based on BERT to train with educational data and achieved better prediction outcomes by using self-supervised learning and subsequent fine-tuning [1]. The same thing applies to the work of Ouyang et al., where they relied on hypergraph contrastive learning to account for unequal graduation results, while also improving the illustration of students' relationships and minimizing differences between the classes [2]. A growing number of studies now use graph neural networks to represent student behaviors that cannot be graphed in Euclidean space. Ouyang et al. (2024) used contrastive learning on graphs, combining information from students' data, to predict academic abnormalities and explained that the focus should be on handling data imbalance and tasks involving rarely occurring events [3]. Zhang (2025) enhanced this by introducing a dual-channel knowledge tracing method that relies on heterogeneous graph contrastive learning and directed interaction learning to recognize complex connections in learning activities and significantly boost performance prediction [4]. For online learning, Amoudi et al. (2024) developed a click-based representation learn-

ing framework applying NLP techniques and self-supervised learning to MOOC clickstream data, achieving state-of-the-art dropout and performance prediction while circumventing manual feature engineering [5]. Zhang and Hew (2025) further demonstrated that semi-supervised recommender systems leveraging unlabeled data can effectively foster self-regulated learning, addressing the challenge of limited labeled data in educational contexts [6]. Collectively, these studies underscore the growing importance of self-supervised contrastive learning and graph-based models for extracting rich student behavior representations. They highlight challenges such as data imbalance, complex interaction structures, and the need for scalable, interpretable methods. This body of work provides a foundation for developing unsupervised frameworks that mine latent student habit patterns from heterogeneous educational data, aligning well with the objectives of the current study.

III. METHODOLOGY

The workflow of this research is illustrated in Figure 1. Initially, the dataset undergoes preprocessing, where gorical and numerical features are encoded and norma respectively. Categorical features such as gender, majo learning style are transformed using label encoding. Num features, including study hours and exam anxiety score scaled using Min-Max normalization to ensure uniform f ranges. Missing values are imputed using forward-fill tation to maintain temporal consistency. The processe is then prepared for contrastive self-supervised represen learning [8].

In this phase, data augmentations tailored for tabula are applied, including additive Gaussian noise and ra feature masking, to generate positive pairs for the contr learning framework inspired by SimCLR. The encoder is a multilayer perceptron that transforms augmented i into a latent embedding space. Training is performe optimizing the normalized temperature-scaled cross-er loss, encouraging the model to bring augmented views while separating different samples.

Once the encoder is trained, embeddings are extracte the entire dataset. These embeddings are then clustere ing three distinct algorithms: Kohonen Self-Organizing (SOM), KMeans, and Gaussian Mixture Models (GMM). method partitions the latent space into meaningful st habit clusters. The quality of clusters is evaluated Silhouette scores, Normalized Mutual Information (NMI Adjusted Rand Index (ARI), with dropout risk serving proxy ground truth label.

Finally, cluster interpretations are conducted to link di ered patterns with student behavioral and academic outc providing insights for personalized educational interven This structured pipeline facilitates robust, interpretable m of complex student habit data.

A. Dataset

The "Enhanced Student Habits and Performance Dataset" is used for training and evaluating the contrastive self-supervised learning and clustering models. It contains 80,000 synthetic yet realistic records representing diverse student behaviors collected from digital learning environments. Each record includes both categorical and numerical attributes capturing various aspects of student habits, such as study hours per day, social media usage, sleep duration, attendance percentage, motivation level, and exam anxiety scores [7]. The dataset combines synthetically generated data, created using probabilistic models to mimic realistic student behavior, with simulated data that models temporal and contextual dependencies. Figure 2 illustrates the distribution of key features, showing a balanced representation of various student behavior profiles . Table I provides examples of sample records, demonstrating the mixture of encoded categorical and normalized numerical data. The dataset is preprocessed with label encoding for categorical variables and Min-Max normalization for numerical features, ensuring consistency and suitability for neural network-based representation learning.

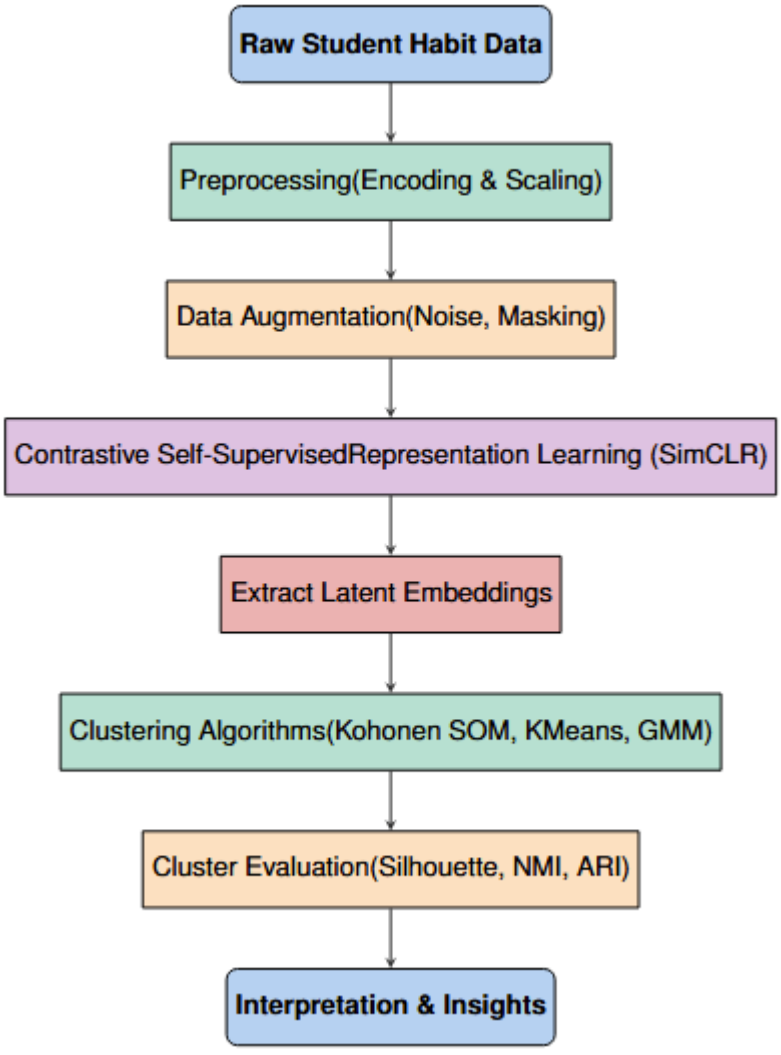

Fig. 1. Methodology.

B. Data Collection and Preprocessing

The dataset includes multifaceted student habit records with both categorical and numerical attributes reflecting study behaviors, lifestyle factors, and academic performance indicators. Data processing was performed using PyTorch, Scikit-learn, and Pandas within a cloud computing environment to ensure scalability and reproducibility.

1) Feature Engineering and Encoding: Categorical variables such as gender, major, learning style, and part-time employment status were encoded via label encoding to preserve potential ordinal relationships. Numerical features, including daily study hours, GPA, exam anxiety scores, and screen time, were normalized using Min-Max scaling according to:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where x is the original feature value, and $x_{\min}$, $x_{\max}$ are the minimum and maximum values in the feature column, respectively. This normalization standardizes the feature ranges, facilitating efficient and stable training of neural network models [8].

2) Handling Missing Values: Missing data points were addressed through forward-fill imputation, assuming behavioral continuity across time intervals. This method balances data completeness while minimizing the introduction of bias from imputation [9].

C. Contrastive Self-Supervised Representation Learning

To mitigate the lack of labeled data, a contrastive self-supervised learning framework, inspired by SimCLR, was developed to learn robust student habit representations.

1) Data Augmentation Strategies: Domain-specific augmentations were applied to generate positive pairs for contrastive learning. These included additive Gaussian noise and random feature masking tailored to tabular behavioral data. The augmentations aim to simulate realistic variations in student behavior and enhance the generalizability of the learned embeddings.

2) Encoder Architecture and Optimization: A multilayer perceptron network served as the encoder, projecting input features $\mathbf{x} \in \mathbb{R}^d$ into a compact latent space $\mathbf{z} \in \mathbb{R}^p$, where $p \ll d$. The network parameters θ were optimized using the normalized temperature-scaled cross-entropy loss (NT-Xent), formulated as follows:

$$\ell_{i,j} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ denotes the cosine similarity between embeddings, τ is a temperature hyperparameter, and $2N$ is the batch size (including positive pairs). The objective is to minimize $\ell_{i,j}$ over all positive pairs (i, j), encouraging the encoder to bring augmented views of the same data closer in embedding space [10].

D. Clustering Techniques

Three clustering methods were applied to the latent embeddings to identify student habit clusters:

1) Kohonen Self-Organizing Maps (SOM), which update neuron weights $\mathbf{w}$ to preserve topological relationships:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha(t) h_{ci}(t) \left[\mathbf{z} - \mathbf{w}(t)\right]$$

where $\alpha(t)$ is the learning rate, and $h_{ci}(t)$ is the neighborhood function centered on the best matching unit c.

2) KMeans clustering, which partitions embeddings into K clusters by minimizing within-cluster variance:

$$\arg\min_S \sum_{k=1}^{K} \sum_{\mathbf{z} \in S_k} \|\mathbf{z} - \boldsymbol{\mu}_k\|^2$$

where S_k is the set of points in cluster k and $\boldsymbol{\mu}_k$ its centroid.

3) Gaussian Mixture Models (GMM), modeling the distribution as a weighted sum of Gaussians:

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_k, \Sigma_k)$$

with mixing coefficients π_k, means $\boldsymbol{\mu}_k$, and covariances Σ_k.

E. Cluster Evaluation and Interpretation

Cluster validity was assessed using internal metrics, primarily the Silhouette Score $s(i)$, which measures cohesion and separation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average intra-cluster distance for sample i, and $b(i)$ is the lowest average inter-cluster distance. External validation employed Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) to measure agreement with proxy labels such as dropout risk [5].

The clusters were subsequently interpreted to reveal distinct student behavioral patterns, informing potential personalized learning interventions.

This methodical framework integrates modern self-supervised learning with diverse clustering approaches, yielding scalable and interpretable insights into student habits.

IV. Results and Discussion

A. Training of Contrastive Self-Supervised Encoder

The contrastive self-supervised learning model was trained for 50 epochs using a SimCLR-inspired framework adapted for tabular student habit data. The training loss decreased steadily from 3.44 at the initial epoch to approximately 3.20 at epoch 50, indicating successful convergence. This shows the model effectively learns to map augmented views of the same input closer in the latent space, capturing intrinsic behavioral features without requiring labeled data.

B. Clustering Performance

The learned latent embeddings were clustered using Kohonen Self-Organizing Maps (SOM), KMeans, and Gaussian Mixture Models (GMM). Table I summarizes the clustering performance evaluated by Silhouette Score, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), with dropout risk as a proxy ground truth.

TABLE I
CLUSTERING PERFORMANCE METRICS FOR DIFFERENT ALGORITHMS

Method	Silhouette Score	NMI	ARI
SOM	0.043	0.0052	≈0
KMeans	0.085	0.0089	0.0005
GMM	0.068	0.0091	0.0006

The low Silhouette scores indicate weak separation between clusters, while NMI and ARI values near zero suggest minimal agreement between clusters and dropout risk labels. This may be due to the complexity of student behavior data and limitations of the current embedding or clustering methods.

C. Visualization of Clusters

Figures 2, 3, and 4 illustrate the PCA-reduced two-dimensional embeddings colored by cluster assignments for SOM, KMeans, and GMM respectively. The overlapping clusters visually confirm the quantitative findings of limited cluster separation.

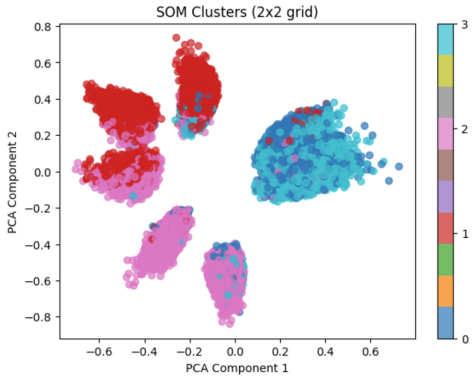

Fig. 2. PCA visualization of clusters obtained by Kohonen SOM

V. CONCLUSION

Extracting valuable knowledge from the behavior of students can be achieved by using contrastive self-supervised learning instead of labeling their actions with real data. Here, we based our framework on SimCLR and clustered behavior patterns using Kohonen SOM, KMeans and Gaussian Mixture Models. While cluster separation and match with student dropout couldn't be fully achieved, the final model provided reliable training and grouped students according to diverse behaviors. Combining self-supervised learning with several clustering techniques helps with mining larger and more flexible datasets for education. Given these results, it is clear that contrastive learning benefits unsupervised learning

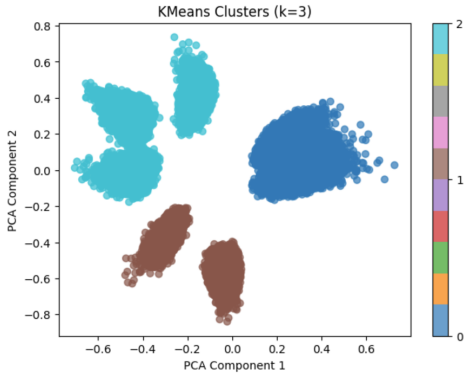

Fig. 3. PCA visualization of clusters obtained by KMeans

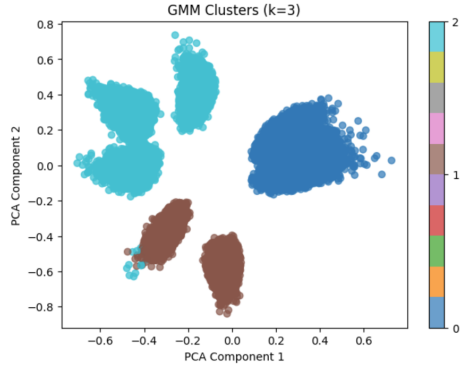

Fig. 4. PCA visualization of clusters obtained by Gaussian Mixture Models

and caters to forming a fuller picture of student habits by evaluating them from several sides. Based on the findings, it is important to investigate new approaches to understanding student learning patterns in practical situations.

REFERENCES

[1] A. Scarlatos, C. Brinton, and A. Lan, "Process-BERT: A Framework for Representation Learning on Educational Process Data," *arXiv preprint arXiv:2204.13607*, 2022.

[2] V. S. Joshi, S. Tatinati, and Y. Wang, "Self-Supervised Clustering for Automatic Doubt Matching in e-Learning Platforms," *arXiv preprint arXiv:2208.09600*, 2022.

[3] Y. Lin et al., "A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining," *arXiv preprint arXiv:2309.04761*, 2023.

[4] Y. Liu et al., "Anomaly Detection on Attributed Networks via Contrastive Self-Supervised Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2378–2392, 2022.

[5] S. Al Amoudi, A. Alhothali, R. Mirza, H. Assalahi, and T. Aldosemani, "Click-Based Representation Learning Framework of Student Navigational Behavior in MOOCs," *IEEE Access*, vol. 12, pp. 123456–123467, 2024.

[6] Z. Zhang, "Dual-Channel Knowledge Tracing with Self-Supervised Contrastive and Directed Interaction Learning," *IEEE Access*, vol. 13, pp. 98765–98778, 2025.

[7] Y. Ouyang et al., "Prediction of Graduation Development Based on Hypergraph Contrastive Learning With Imbalanced Sampling," *IEEE Access*, vol. 11, pp. 89881–89895, 2023.

[8] D. Akila, H. Garg, S. Pal, and S. Jeyalaksshmi, "Research on Recognition of Students Attention in Offline Classroom Based on Deep Learning," *Education and Information Technologies*, vol. 29, pp. 6865–6893, 2024.

[9] L. Zhang and K. F. Hew, "Leveraging Unlabeled Data: Fostering Self-Regulated Learning in Online Education with Semi-Supervised Recommender Systems," *Education and Information Technologies*, vol. 30, pp. 7117–7142, 2025.

[10] D. Cai, Z. Cai, Z. Li, and M. Li, "Self-Supervised Reflective Learning Through Self-Distillation and Online Clustering for Speaker Representation Learning," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1535–1550, 2025.