IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE EDUCATION SOCIETY SECTION

Received 27 July 2024, accepted 17 August 2024, date of publication 27 August 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450514

RESEARCH ARTICLE

Click-Based Representation Learning Framework of Student Navigational Behavior in MOOCs

SHROOQ AL AMOUDI[1], AREEJ ALHOTHALI[1], RSHA MIRZA[1], HUSSEIN ASSALAHI[2], AND TAHANI ALDOSEMANI[3]

[1]Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]English Language Institute, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3]College of Education, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia

Corresponding author: Shrooq Al Amoudi (sabdullahalamoudi0001@stu.kau.edu.sa)

This work was supported by the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, through project number IFPRC-039-126-2020, and by the Deanship of Scientific Research at King Abdulaziz University, Jeddah, Saudi Arabia.

ABSTRACT Predictive learning outcomes' models for online students can provide useful information to instructors to estimate students' final performance in the early stages of a course. Anticipating student performance can improve learning efficiency. Existing research models that analysed student data have focused on handcrafted features, but these models have limitations in exploring new behavioral patterns that indicate student performance and how they can be used in online courses. The clickstream data contains a significant amount of information that accurately describes students' learning processes, which makes it difficult to construct using hand-crafted features. To analyze student behavior effectively, we attempted to capture critical knowledge from the field of natural language processing (NLP) to the field of student performance prediction in Massive Open Online Courses (MOOCs), owing to how closely they resemble each other. In this article, we propose a novel framework for automatically producing useful data representation that enhances prediction outcomes using student learning behavior clickstream data with a self-supervised learning approach. First, we developed a self-supervised clickstream pre-training setup to model learner click generation. Second, we adjusted these latent representations before applying them to a downstream supervised learning task. Extensive experimental results on two real-world datasets demonstrated that the proposed approach is effective. The combined approach of skip-gram embeddings with Principal Component Analysis (PCA) achieved the highest accuracy, particularly on the Xutangx dataset, with an accuracy of approximately 72.70% and an F1-score of approximately 81.03%. Furthermore, when applied to the KDDCUP dataset, this methodology exhibited even higher performance, with an accuracy of 80.91% and an F1-score of 87.42%. Our results showed the potential of NLP techniques to improve dropout prediction in MOOCs by extracting informative representations from clickstream data, allowing a deeper understanding of student behavior, and facilitating early intervention strategies.

INDEX TERMS Student modeling, representation learning, self-supervised learning, skip-gram model, clickstream data, MOOC.

I. INTRODUCTION

Online learning platforms and Massive Open Online Courses (MOOCs) such as Coursera, edX, and Udemy have recently become popular due to the expansion of extensive Internet-based educational materials [1]. These platforms provide high-quality educational materials, such as lecture videos, assignments, and quizzes from multiple elite schools [2].

Research suggests that compared to face-to-face learning, online learning platforms have a positive impact on education quality, and learners' participation has increased significantly [3]. Driven by this, since 2012, interest in MOOCs has surged to address several challenges to learning, including

The associate editor coordinating the review of this manuscript and approving it for publication was Nkaepe Olaniyi.

financial, geographical, and recently global pandemic. With the advent of the Covid-19 pandemic in 2020, schools worldwide switched to online learning through online learning platforms instead of traditional face-to-face learning [4]. The Udemy MOOCs platform released a report showing an increase in global online learning. The first month of the start of the pandemic witnessed a 425% spike in individual enrollment, a 55% increase in the creation of new courses and the use by businesses and government of online platforms has rocketed to nearly 80% [5].

Despite the widespread use and usefulness of Moocs in mass education, several writers have highlighted the challenges of quality learning interventions in online learning environments [6]. To address this gap and enhance students' online learning engagement, research highlights the need to provide innovative approaches to collect real-time data to detect learning behaviors and inform pedagogical approaches to learning [7]. This can be attained by collecting data pertaining to the dynamic online learning activities, including watching lectures, taking quizzes, participating in discussion forums, and interacting with course materials. These activities provide a thick pool of data that can be gathered and stored on the MOOCs platform [3], [8], [9]. Data mining techniques are then utilized to analyze key patterns in students' behavior collected in log data and to provide useful pedagogical insights for instructors and students [8], [9], [10].

Educational Data Mining (EDM) is an emerging multidisciplinary research topic that applies statistics and machine learning to education. Major data mining techniques include regression, classification, and clustering. These algorithms can be applied to massive educational datasets to identify patterns and inform pedagogical intervention. Research suggests that EDM effectively captures massive data within MOOCs and provides deeper insights into how to improve online teaching methods and inform pedagogical approaches [11], [12].

Student modeling is a crucial field of research in EDM for generating predictions of personal attributes that influence student learning outcomes. Personalized models based on developing predictive models of student learning behaviors including performance or engagement indicators can considerably improve the efficacy of learning in MOOCs [13], [14]. This process starts by collecting a series of student behavioral clickstream data, which is a time-stamped log record of click events and provides fine-grained information regarding student learning. This type of data allows researchers and instructors to gather information on the way each student navigates and interacts with online educational materials, potentially providing rich insight into the student's learning processes and determining behavioral patterns relevant to student learning outcomes. However, due to the vast array of student clickstream data, manual analysis of large logs is extremely difficult to achieve. In addition, human labeling is both time-consuming and costly. Therefore, supervised learning approaches are considered impractical [15].

In this context, there is growing interest in developing frameworks to analyze online student data through machine learning and deep learning methods. Several approaches implemented in this body of research to predict student performance include predicting student performance based on grade prediction [7], [14], [16], predicting student performance based on a student's at-risk prediction [17], and other studies predicting student performance based on students' failure or success prediction [18], [19]. Moreover, some research has been conducted on dropout, which is one of the most popular challenges faced by educational institutions [20], [21]. Previous research sought to predict and improve student performance in MOOCs by considering student behavioral data with demographic data [16], [17], [22], problem-solving clickstream data [7], [14], [19], video-watching behavioral data [20], [23], and course data [17]. Most studies extract features by hand, feature engineering-based models, or depend on labeled datasets for predicting student learning outcomes. Brinton [14], presented a personalized prediction model for predicting students' correct first attempt (CFA) grades from student assessment and video-watching behavioral data. Rahman and Islam [16], applied different classification algorithms such as Naïve Bayes, Artificial Neural Network (ANN), decision tree, and K-Nearest Neighbor (K-NN) to show how the behavioral and absence features of the student impact academic performance. Adnan et al. [17], developed a predictive model for identifying at-risk dropout students at different times of the course for early intervention using several classification algorithms based on demographics and assessment scores. Recent research has focused on analyzing temporal features gained from raw clickstream data with deep learning models to enhance the prediction of student' outcomes [13], [23], [24].

A major drawback of the research conducted in this field pertains to extracting features by hand, developing models that cannot be used for real-world educational scenarios, and not using all the raw clickstream data generated from MOOC platforms. Furthermore, a significant obstacle in analyzing educational process data is determining the method of its representation. As a result, existing approaches for representing process data tend to be tailored to specific forms of educational process data such as video-watching clickstreams [23] and problem-solving clickstreams [24]. The challenge of effectively representing process data is emphasized by the fact that superior representations yield enhanced performance in subsequent tasks, as opposed to relying on the conventional approach of one-hot encoding, which is prevalent in many existing models for representing educational process data. Thus, the motivation of this study is to build a bespoke framework capable of addressing three main issues. Specifically, it sought to accurately identify student behavior patterns by overcoming the need for a large number of labels in a supervised learning approach. The framework also eliminates the requirement for manual feature

engineering, utilizing all student data processes that influence the students' performance when interacting with the platform. In addition, it automatically learns useful representations of educational process data that can be adjusted across many different real-time learning scenarios on target prediction tasks.

Based on the foregoing discussion, this study aims to answer the following research questions:

1) How can the developed framework reveal the analytics of the correlation between student navigational behavior in MOOCs and dropout prediction?
2) Can representation of all clickstream data generated by the platform improve prediction accuracy?
3) Does clickstream data representation produce meaningful clusters of student behavior patterns?

To address these questions, we attempted to capture the knowledge gained from the Natural Language Processing (NLP) field, which enables the representation of words as numbers called One-Hot encoding. However, it turns out that the embedding models perform better if each word is represented by trainable vectors with a multidimensional space. Word embedding improves the representation ability of the word vector, thus improving the model performance. Based on this, the fundamental concepts of developed framework were derived from NLP research and modified for student learning behavioral data. We developed a novel click-based representation learning framework for student navigational behavior in MOOCs to predict student learning outcomes and assist in eliminating early dropouts. IN particular, the genuine contribution of the proposed model was to solve the feature learning problem by utilizing all clickstream data generated by Mooc's platform via a self-supervised learning approach to learn a good latent representation of educational process data with a large degree of redundancy. A major technique adopted to improve the prediction quality is the use of a self-supervised learning approach for handling data lacking labels and automatically capturing student sequences. These features can potentially lead to the production of meaningful clusters of student representations.

The contribution of this research is threefold:

- The study presents a novel two-step process for representation learning on educational process data, designed to enhance the predictive capabilities of the model. First, a self-supervised natural language model is developed to generate dense, continuous embeddings (GloVe and Skip-gram) of learner clickstream data. These latent representations are then refined (using PCA) before being applied to downstream supervised learning tasks to predict learners' outcomes.
- The proposed approach leverages all types of clickstream data sequences (e.g., video-based, forum-based, problem-based data) generated from e-learning platforms to predict learners' outcomes. By not eliminating any clickstream type, this approach captures a comprehensive view of students' learning behavior, thereby improving the accuracy of predictive tasks.
- The study also provides an analysis of learners' clusters based on the learned representations of their clickstream data. This analysis can help in delivering personalized interventions or adaptive learning strategies tailored to students' characteristics.

II. RELATED WORK

In this section, we describe different feature extraction methods for predicting dropout using clickstream data sequences.

- Feature Engineering-based Data Representation.
  Feature engineering is the process of selecting and transforming relevant variables from the raw data when building a predictive model. The purpose of feature engineering is to improve the performance of machine learning algorithms. In feature engineering, a classifier is trained by using a set of manually defined features. This technique has been successful for different forms of educational data [24]. Nevertheless, defining and implementing these features require a significant amount of engineering work based on domain knowledge. Thus, feature engineering is time-consuming, not very precise, and not generalizable [25]. To predict student learning outcomes, the majority of studies either manually extract features, use feature engineering-based models, or rely on labeled datasets (annotations) [14], [16], [17]. Similarly, Imran et al. [26], proposed a feed-forward artificial neural network model to predict student dropout based on students' behavioral data, using the HMedx dataset. Xiong et al. [27], presented a Recurrent Neural Networks (RNN) with a Long Short-Term Memory (LSTM) network as an RNN-LSTM model that utilized statistical weekly features collected from student log data to predict learner dropouts on the Chinese university MOOCs platform. Yin et al. [28], utilized the KDDcup dataset to develop a model that combined the self-attention mechanism and conditional random field (CRF) for dropout prediction in MOOCs. On the other hand, some studies applied feature selection techniques for automatic dimensionality reduction and selected the key features that affect prediction tasks such as Principal Component Analysis (PCA) [7]. Most of these models, which rely on handcrafted features have operational limitations regarding online courses. For example, if a feature is defined for specific course exercises, it cannot be used in a subsequent offering if the exercises are removed. Moreover, it is difficult to design effective handcrafted features that can characterize students' learning behaviors such as engagement or performance indicators for prediction tasks.
- Deep Learning-based Data Representation.
  Deep learning feature extraction involves employing pre-trained deep neural networks to automatically derive meaningful features from raw data, which can include

images, text, or other forms of high-dimensional data. Specifically, Convolutional Neural Networks (CNNs) are commonly utilized for image data, whereas RNNs are employed for sequential data such as text. These deep learning models can identify complex patterns and representations. Recent research has focused on analyzing some statistical and temporal features with deep learning models from raw clickstream data to enhance the prediction of student outcomes [13], [23], [24]. To predict dropout and extract features from statistical temporal features, several studies have built CNN. Qiu et al. [29], employed the KDDcup dataset to develop a two-dimensional CNN (DP-CNN) model to analyze the learning patterns of students as day-to-day statistical behavioral features over a period of time were encoded in a matrix. Wang et al. [30], developed an automatic feature extraction dropout prediction model using raw data. To build matrices that describe a student's learning behavior over a couple of days, the raw data from each activity record were first encoded into a one-hot vector and concatenated. Subsequently, time-series relationships were utilized and features were extracted using a CNN and RNN model. Similarly, Zheng et al. [21], Developed the FWTS-CNN model to address the limitations of traditional CNNs in dropout prediction tasks, which often overlooks student behavioral features and time series effects. The FWTS-CNN extracts statistical behavioral features from students' learning activities, uses a decision tree for feature weighting and selection, and generates a time series matrix. This matrix is then fed into the CNN model to predict student dropout in the MOOCs, utilizing the KDDcup dataset. Mubarak et al. [31], utilized the CAROL and KDDcup datasets to develop a CONV-LSTM hyper-model that combines long short-term memory with CNN to automatically extract features from raw data for predicting student dropout. A cost-sensitive method was used for the loss function to improve prediction performance. Wu et al. [32], proposed the CLMS-Net model to overcome the limitations of handcrafted feature extraction methods. They aggregated CNN, LSTM and Support Vector Machine (SVM) for the daily prediction of student dropout, utilizing the KDDcup dataset. Furthermore, Yu et al. [18], identified seven types of cognitive participation models of students by establishing a series of learning behaviors utilizing video clickstream records of students from the KDDcup dataset to overcome the problem of low completion rates in MOOCs. They used N-gram feature extraction methods and employed (K-NN), (SVM) and (ANN) algorithms to predict whether learners pass the course based on their learning patterns. Additionally, they determined which types of videos had the greatest impact on whether students passed the course and made this information available to instructors for reference. Moreover, Ding et al. [20], employed edx dataset to address the issue that when models are transferred, they automated transductive transfer learning to improve prediction performance when models are transferred between courses. By organizing MOOC video clickstream data temporally and categorizing events, they proposed using auto-encoders to compress the data into latent space representations. Two methods were developed: Passive-AE Transfer, using transductive PCA, and Active-AE Transfer, using a correlation alignment loss term. The results show that Passive-AE Transfer excels in transferring between different courses, whereas Active-AE Transfer is better for transferring between different offerings of the same course.

Previous studies on student dropout prediction have used different data representation approaches. The majority of previous research utilized statistical techniques to analyze raw clickstream data, whereas others employed basic one-hot encoding to prepare raw data for deep learning models. Wu et al. [32], proposed that the CLMS-Net model improves student dropout prediction by using CNN, LSTM, and SVM to extract features from raw data automatically. It encodes data into a one-hot vector and represents the daily student activity in matrices. CLMS-Net outperforms traditional feature engineering models and other neural networks such as GRU-RNN. Similarly, Zhang et al. [33] proposed a hybrid deep neural network model to predict student dropout patterns using Convolutional Neural Networks (CNN) and Squeeze-and-Excitation Networks (SE-Net) to extract learner behavior on an online platform. They analyzed the time series relationship between learning behaviors using a Gated Recurrent Unit (GRU) network. They used one-hot encoding to transform clickstream data into a vector with hybrid CNN-RNN and CNN-SE-GRU. However, little if any study focused on encoding all learner clickstream data in a self-supervised manner. In terms of student interactions related to personal knowledge mastery, there is a large degree of information redundancy in the raw clickstream data, thus discarding some of a student's behavioral clickstream data or relying on statistical features limits the ability of machine learning to explore new behavioral patterns that indicate students' performance. No study has utilized all behavioral clickstream data, particularly with the self-supervised learning approach. Thus, our model automatically extracts features to provide useful data representation by deeply capturing student learning behavioral data which assists in improving the performance of the real-time prediction tasks. Our framework is distinguished using a self-supervised learning approach with all temporal raw clickstream data generated by the MOOCs platform. Finally, the framework uses classification and clustering methods to identify student behavioral patterns.

III. PROPOSED METHOD

This section outlines the Modeling Strategy, elucidates the fundamental concepts of Data Representation,

and presents an overview of our dropout prediction framework.

A. MODELING STRATEGY

The main components of the framework used in our modeling strategy are illustrated in Figure1, highlighting the key processes and methodologies implemented for effective dropout prediction in MOOCs. The first step in model development was to represent the logs of student interactions with the platform to predict whether students would drop out of a course based on their early interactions with the platform. This involves extracting meaningful information from the raw data and transforming it into a format that a machine-learning model can use. One approach to processing data is to use skip-gram word embeddings to represent the student logs. Skip-gram is a word embedding algorithm that learns the relationships between words in a corpus. This can represent the sequence of actions that a student takes as a vector of numbers. We can then use this vector to represent each student's log. An n-dimensional vector is generated for each student, which can then be used as an input to a machine-learning model. Finally, the performance of the developed model was evaluated using various metrics, such as accuracy, precision, recall, and F1 score.

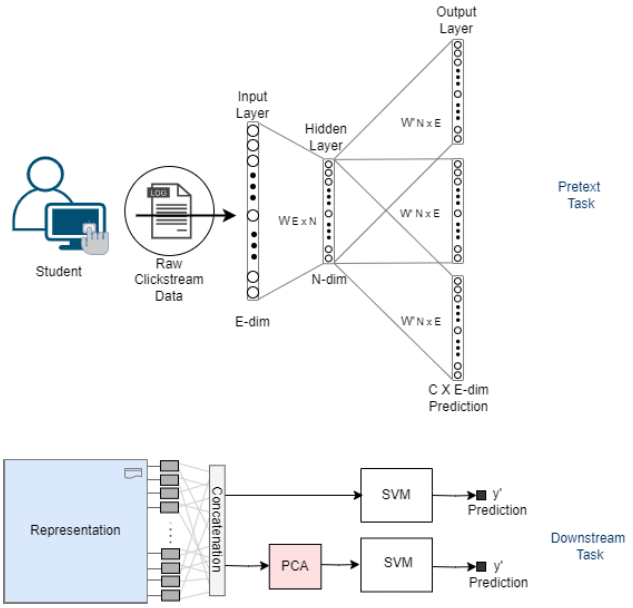

FIGURE 1. The proposed framework for learning meaningful representations of student actions through a two-step process. First, leveraging clickstream data, the pretext task generates initial action representations, which our model then refines (downstream task).

B. LEARNING DATE REPRESENTATION

Instead of spending a significant amount of time manually developing effective features, deep learning enables the acquisition of new efficient feature representations from the training data [34]. The machine learning community has recently concentrated its efforts on representation learning, which converts real-world data such as graphs, pictures, and texts into representations that machine learning algorithms can use efficiently [25]. As a novel feature extraction method, deep learning has made enormous in the field of progress in data mining. Traditional machine learning methods often concentrate on manually extracted, handcrafted feature representations. To ensure that the model produces invariant and untangled results and to improve accuracy and performance, representation learning is required. This indicates that deep learning uses features from massive datasets rather than handcrafted features which heavily rely on designer knowledge and cannot be completely utilized with big data. Deep learning models can be automatically trained using large datasets with millions of parameters to represent the features of those datasets [34]. Two approaches are used in data representation. First, the concept of one-hot encoding, which is used as a baseline model to represent categorical data in machine learning models, is explored. Next, the skip-gram algorithm, which is popular in natural language processing tasks, particularly for word embedding, is discussed. Finally, we examined PCA, a statistical technique used for dimensionality reduction, to extract meaningful information from complex datasets.

1) ONE-HOT ENCODING

One-hot encoding is an approach for depicting categorical data in which each distinct category or class is denoted by a distinct binary value. In this encoding method, every category is assigned a binary code, with only one bit marked as ''hot'' or set to 1, whereas the others are marked as ''cold''. One-hot encoding is a simple technique in which each unique action in the click-stream data is converted into a binary vector of length equal to the number of exceptional actions. In this vector, all elements are set to zero except the element corresponding to a specific action, which is set to one. This results in a sparse matrix representation in which each row corresponds to a single action, and each column corresponds to a unique action in the dataset. Despite its simplicity, one-hot encoding ensures that each action is distinctly represented without implying any inherent relationship between different actions.

2) SKIP-GRAM MODEL

The skip-gram model [35], utilized in natural language processing and self-supervised learning, belongs to the category of word embedding models. Skip-gram belongs to the Word2Vec algorithm family and is intended to learn distributed word representations in a continuous vector space. The core concept of the skip-gram model is to predict the context words for a specified target word within a given text window. Through extensive training on substantial text corpora, the model gains the ability to represent words in a manner that captures their semantic relationships, thus enabling a sophisticated understanding of word meanings based on their contextual usage. This approach has found widespread application in various tasks, including language modeling, information retrieval, and word similarity analysis.

The skip-gram model is a simple feedforward neural network with three layers, input, hidden, and output, designed to learn distributed representations of user actions in clickstream data. It starts by defining the vocabulary of unique actions in the click-stream data, where each action is represented as a one-hot encoded vector. The model is trained by adjusting the weights to minimize the difference between the predicted probabilities and the actual occurrences of actions in the sequences. These embeddings capture both semantic and syntactic relationships, positioning semantically similar actions in close proximity within the resulting vector space.

In the skip-gram model, the vector representation of the input event E_I is determined as follows:

$$v_{E_I} = W^T \delta(E_I)$$

where W^T denotes the transposed weight matrix on the left side and $\delta(E_I)$ represents the one-hot encoding of the input event. Hidden layer h is computed as follows:

$$h = W \cdot \delta(E_I)$$

where W is the weight matrix of size $|V| \times d$. A softmax layer, commonly used in classification tasks, generates a probability distribution across events, aiming to predict events in context. The probability of an output event E_O, given an input event E_I, is calculated as:

$$p(E_O \mid E_I) = \frac{\exp(W' \delta(E_O) \cdot v_{E_I})}{\sum_{j=1}^{V} \exp(W' \delta(E_j) \cdot v_{E_I})}$$

where W' denotes the weight matrix of the right side, and $\delta(E_O)$ represents the one-hot encoding of the output event. This calculation involves exponentiating the dot product of the input vector v_{E_I} and the output vector $W' \delta(E_O)$, and then normalizing the result by the sum of the exponentiations of all possible events' vectors.

The training objective of the skip-gram model is to minimize the negative log-likelihood of the context actions given the target action. This is achieved through cross-entropy loss, which measures the discrepancy between the predicted and actual event sequences for all students. Cross-entropy loss C is computed as follows:

$$\mathrm{C} = -\sum_{s \in S} \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le i \le c, i \neq 0} \log p(E_{t+i} \mid E_t)$$

where c represents the context window size, S is the set of all the students, and T is the length of the sequence for each student. This loss is used for backpropagation to update the model parameters, ensuring that the weights are adjusted to accurately capture the relationships between actions in the clickstream data. After training, the rows of the weight matrix W serve as the learned embeddings E of the actions. These embeddings effectively capture the context and relationships between actions, making them suitable for further analyses and applications.

3) PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) [36] is an unsupervised learning technique, a commonly employed method in machine learning and statistics for reducing dimensionality. Its primary objective is to convert high-dimensional data into a more compact representation while preserving the crucial variance within the dataset. This is accomplished by identifying and highlighting the principal components, which are the linear combinations of the original features. PCA helps to reduce a complex dataset to a lower-dimensional space, also called a latent space, thus revealing hidden dynamics. Additionally, it helps to obtain the most meaningful basis for re-expressing any noisy dataset. After obtaining the embeddings from the skip-gram model, PCA was applied to reduce their dimensionality.

The first step in PCA is to center the data by subtracting the mean embedding vector μ from each embedding. This ensures that the data have a mean value of zero:

$$E' = E - \mu$$

where

$$\mu = \frac{1}{n} \sum_{i=1}^{n} E_i$$

Here, E represents the matrix of embeddings, and n is the number of embeddings.

The centered data E' are then used to calculate the covariance matrix C, which captures the variance and relationships between the different dimensions of the embeddings:

$$C = \frac{1}{n} E'^T E'$$

To determine the principal components, we perform an eigen value decomposition of the covariance matrix C. This involves solving for the eigenvalues and corresponding eigenvectors as follows:

$$Cv_i = \lambda_i v_i$$

where λ_i and v_i are the eigenvalues and corresponding eigenvectors respectively. The eigenvectors represent the directions of maximum variance, and the eigenvalues indicate the magnitude of variance in these directions.

The top k eigenvectors corresponding to the largest k eigenvalues are selected to form the projection matrix P. This step reduces the dimensionality of the data, while preserving as much variance as possible:

$$P = [v_1, v_2, \ldots, v_k]$$

Finally, the centered embeddings are projected onto the new subspace defined by the principal components. This transformation yields the following PCA-transformed embeddings:

$$E_{PCA} = E'P$$

The PCA-transformed embeddings E_{PCA} reduced dimensionality, retaining the most significant variance in the data.

These transformed embeddings can now be used for various downstream tasks, such as clustering, and visualization, or as features in predictive models. By combining, the skip-gram model for embedding generation with PCA for dimensionality reduction, the methodology effectively captures and simplifies complex behavioral data, thereby improving its readability for interpretation and analysis.

C. DROPOUT PREDICTION

The problem with dropout prediction is a sequence prediction task designed to forecast a student's future course attendance by examining their activity records over time. Researchers have different definitions of student dropout in MOOCs; some studies define dropout based only on learning behaviors [23], [37], while others define dropout depending on whether certification is acquired [38]. The developed framework predicts dropout by tracking a student's activity record during the first 100 actions, considering that our Skip-Gram representation learned in a self-supervised approach to students' learning behaviors is good enough to perform effectively. In this binary prediction scenario, two distinct labels are employed: one is assigned to dropouts, and zero is assigned to non-dropout.

IV. EXPERIMENTS AND DISCUSSION

This section presents experiments performed to validate the impact of the proposed model. First, the dataset used in the framework is introduced. Second, the experimental settings of the proposed model were described. We then show the experimental comparison results of our framework with several methods and verify the validity of our proposed model through both quantitative and qualitative experiments.

A. DATASET

As previously explained, this study aims to develop a framework for automatically learning the representation of student behavior from clickstreams generated from online educational platforms for student dropout prediction. The developed framework was trained and tested using two public activity log datasets from the XuetangX Chinese learning platform [39]. The first dataset is XuetangX, which contains 68M action items of Instructor-Paced Mode courses (IPM) and 48M action items of Self-Paced Mode courses (SPM) from around 378,237 students, where each action item represents accessing a specific course event, such as video navigation, forum navigation, or problem navigation. In this study, approximately 5.5M activity logs were used, containing 43 IPM courses spanning over four months, and around 36,000 learners' data. In each course, records of participation data were provided regarding cumulative student activity. There are 21 distinct actions, classified into five categories including (discussion forum, problem, video, click, and close actions). The learner action categories in the XuetangX dataset are listed in Table 1. Additionally, the XuetangX dataset encompasses learner profile details such as gender, age, location, and education level, along with course category, as well as the dates the course begins and ends. The second dataset is KDDCUP which Contains 39 IPM and 120,542 learner data. In each course, records of participation data were provided regarding cumulative student activity. There are seven distinct actions (navigate, access, problem, page-close, video, discussion, wiki). The learner actions in the KDDCUP dataset are presented in Table 2.

TABLE 1. Learner action categories of the XuetangX dataset.

Action Categories	Action Name	Description
Video Actions	Seek_video	Seeking forward or backward a lecture video.
	Play_video	(Re)playing button in a lecture video.
	Pause_video	Pausing a lecture video.
	Stop_video	Stopping a lecture video.
	Load_video	Loading a lecture video.
Problem Actions	Problem_get	Opening an assignment.
	Problem_check	Checks to verify answers.
	Problem_save	Saves current progress.
	Reset_problem	Resets an assignment.
	Problem_check_correct	Correct answer verified.
	Problem_check_incorrect	Incorrect answer verified.
Forum Actions	Create_thread	Creating a discussion thread.
	Create_comment	Posting a discussion comment.
	Delete_thread	Deleting a discussion thread.
	Delete_comment	Deleting a comment.
Click Actions	Click_info	Visiting the info page.
	Click_courseware	Accessing the educational material.
	Click_about	Visiting the about page.
	Click_forum	Viewing a discussion forum.
	Click_progress	Checking overall progress.
Close Actions	Close_courseware	Leaving educational material.
	Close_forum	Closing the discussion forum page.

TABLE 2. The learner actions of the KDDCUP dataset.

Action Name	Description
Navigate	interacting with elements other than a video.
Access	Accessing elements other than a video.
Problem	Doing assignment.
Page_close	Closing the web page.
Video	Watching a lecture video.
Discussion	Viewing a discussion forum.
Wiki	Reading the Wikipedia of a lecture.

B. EXPERIMENTAL SETTINGS

Feature representation data algorithms were implemented for the raw clickstream data sequences (Section III-B). The scope was confined to the first 100 sequences of actions to examine our model's capacity to predict student dropouts using the baseline model.

For classification and clustering objectives, the developed framework was compared with the baseline model. In the classification problem, we concentrate on two classes of student performance, whereas in the clustering task, the goal is to find distinct groups based on students' behavior patterns.

- Evaluation Criteria
  The evaluation criteria were the benchmarks used to measure the performance and effectiveness of predictive models. In the context provided, these criteria were applied to assess the predictive model performance.

Various evaluation metrics were selected to evaluate the performance of the predictive models:

1) ACCURACY

Accuracy is a fundamental metric that indicates the proportion of correctly predicted instances out of the total instances. In the context of student performance prediction, it reflects the overall correctness of the model, showing how often the model correctly identifies whether a student will drop out or continue. Higher accuracy indicates that the model is generally reliable, but it may not always capture the nuances of imbalanced datasets where one class, such as students who do not dropout, may dominate.

2) PRECISION

Precision measures the proportion of true positive predictions among all instances classified as positive, reflecting the model's ability to avoid false positives. In the context of this study, it reflects the model's ability to correctly predict students who are at risk of dropping out without mistakenly classifying students who are likely to continue as dropouts.

3) RECALL

Recall, also called sensitivity, evaluates the proportion of true positive predictions among all actual positive instances, showcasing the model's capacity to capture relevant instances. For student dropout prediction, recall indicates the model's capacity to identify all students who are truly at risk of dropping out. High dropout recall is crucial because it enables instructors to identify all students at risk of dropping out. This ensures that no potentially at-risk student is overlooked, allowing for comprehensive intervention strategies.

4) F-SCORE

The F-score is a balanced metric that considers both precision and recall, calculated as their harmonic mean. This metric is particularly useful when there is a need to balance between the model's ability to correctly identify true positives, such as students who will drop out, and its ability to avoid false negatives. In student performance prediction, a high F-score suggests that the model is effective in identifying at-risk students while minimizing missed identifications, making it a critical metric in early intervention strategies. Table 3 lists the mathematical equations for the evaluation criteria.

C. QUANTITATIVE EXPERIMENTS

There are two phases in our framework to capture meaningful representations of user interaction. Initially, the pretext task generates basic action representations using the clickstream data. Subsequently, our model enhances and fine-tunes these representations in a downstream task.

TABLE 3. Evaluation criteria and equations.

Evaluation Criteria	Mathematical Equation
Accuracy	$\frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$
Precision	$\frac{\text{True positives}}{\text{True positives + False positives}}$
Recall	$\frac{\text{True positives}}{\text{True positives + False negatives}}$
F-score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

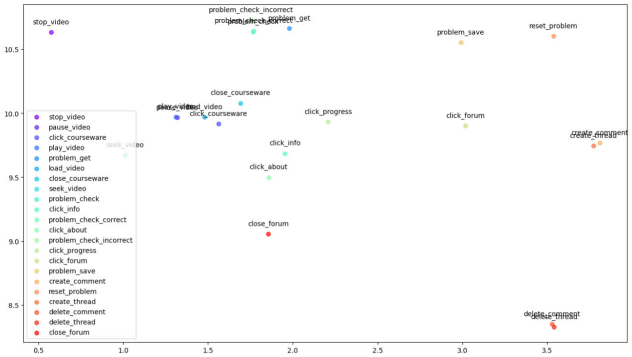

FIGURE 2. XutangX action representations using PCA technique.

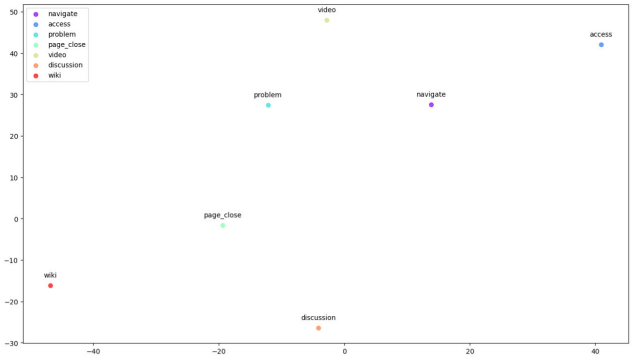

FIGURE 3. KDDCUP action representations using PCA technique.

1) PHASE 1: PRETEXT TASK

A pretext task is known as pre-training, which assists the model in learning useful intermediate data representations. Moreover, it is helpful for a deep understanding of the structural meaning of raw data to be used in practical downstream tasks. The pre-training stage applies a set of predictive tasks to process data to help the model learn several general features through learned representation. In this manner, we find a similarity heatmap for data representation approaches, which is a mathematical construct used to quantify the similarity between pairs of objects or entities within a given dataset. It is typically represented as a square matrix where each cell corresponds to the similarity score between two objects. The similarity score can be computed using various techniques depending on the nature of the data, such as Euclidean distance, cosine similarity, or correlation coefficients. One of the key advantages of using this similarity heatmap

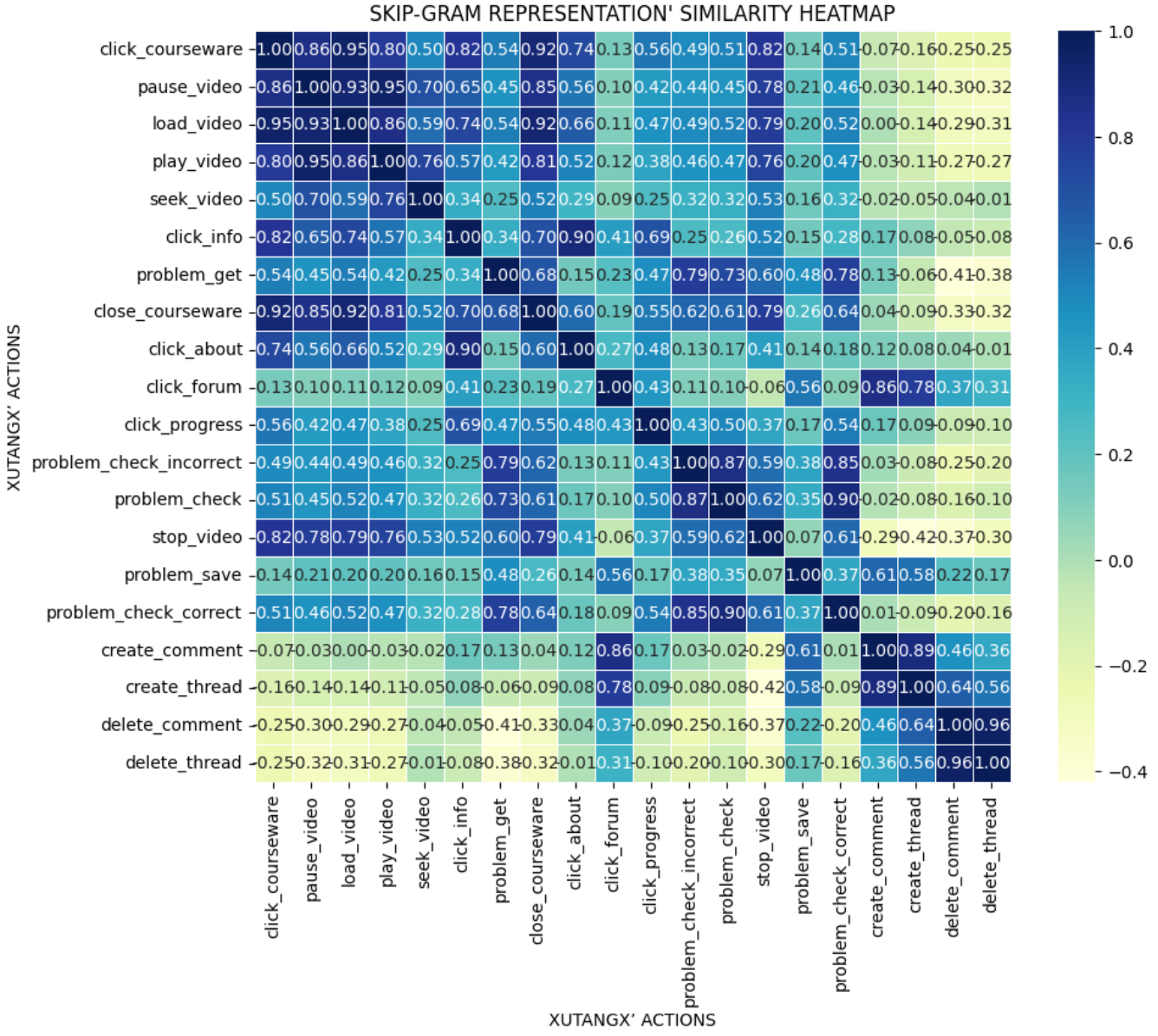

FIGURE 4. Similarity heatmap between XutangX' actions using skip-gram representation.

is its ability to provide a structured representation of the relationships between actions in a dataset. We create the learning behavior navigation space using one-hot encoding for baseline representation (Section III-B1). This process involves generating a 21-dimensional action vector for XutangX and a 7-dimensional action vector for KDDCUP, with a binary indicator for each action sequence. Specifically, the vector contains a 1 if the action is present in the user's sequence and a 0 otherwise.

In contrast, we apply the skip-gram algorithm (Section III-B2) as a deep learning approach within the Word2Vec framework, specifically designed for generating word embeddings. This deep-learning algorithm has proven to be effective in capturing intricate semantic relationships between words. We were motivated by the skip-gram model to represent students' learning behavior navigation. This application of the algorithm is state-of-art, as it is traditionally employed for word sequences rather than behaviors. This model is capable of generating a vector representation of student behavioral actions by analyzing their contextual relationships within extensive daily action sequences as a large corpus. We set the action vector size to 100 and window size to 5. The similarity matrix, derived from the skip-gram representation of the Xutangx student actions in Figure 4, provides valuable insights into the relationships between different actions. High similarity scores between certain actions, such as "click_courseware" and "pause_video," suggest a strong association, indicating that these actions

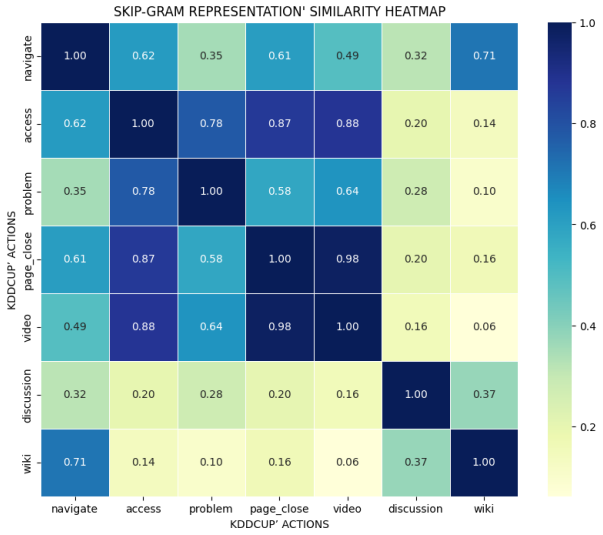

FIGURE 5. Similarity heatmap between KDDCUP' actions using skip-gram representation.

frequently occur together. Moderate similarity scores, like those between ''load_video'' and ''play_video,'' imply functional relationships, possibly indicating sequential actions or related tasks. Conversely, actions with low similarity scores, such as ''click_forum'' and ''click_progress,'' indicate distinct behaviors that are less likely to co-occur. Furthermore, as shown in Figure 5, the similarity heatmap represents the nearest neighbor skip-gram actions representation of the KDDCUP dataset. The similarity matrix from the skip-gram model revealed the relationships between different student actions. High similarity scores between actions, such as accessing and watching videos indicate these behaviors often occur together. Moderate similarities, such as navigating and accessing wiki pages, suggest functional relationships, whereas low similarities, such as discussions and watching videos, highlight distinct behaviors. Overall, the model effectively identified clusters of related actions, which can inform further analysis or interventions to enhance understanding of student interactions with educational content. Moreover, this study used feature dimensionality reduction PCA (Section III-B3) to comprehensively analyze the knowledge states of students, automatically reduce the analysis targets, and minimize the loss of information. We set the number of components to two. Using Matplotlib, a scatter plot was created, to visualize where each point was a word vector and was labeled and colored according to the associated event types. This approach transforms each learned vector representation into two dimensions, ensuring that the proximate points and dissimilar vectors representing similar vectors are represented by points that are likely to be distant. Figure 2 and Figure 3 show the XutangX and KDDCUP action representations, respectively, using the PCA technique. The heatmap of Figure 6 indicates an increase in the similarity between the Xutangx events. For instance, many actions now show near-perfect similarity scores (close to 1) with

TABLE 4. Comparison of performance scores with our framework and the baseline model.

Dataset	Model	Method	Result			
			Accuracy	Precision	Recall	F1-Score
XUTANGX	SVM	One-hot	63.50	79.64	70.23	74.64
		Skip-gram	71.87	86.90	74.17	80.03
		PCA	**72.70**	85.86	76.71	**81.03**
KDDCUP	SVM	One-hot	39.08	96.46	24.27	38.79
		Skip-gram	80.75	92.09	82.83	87.22
		PCA	**80.91**	91.77	83.46	**87.42**

each other, such as ''click_courseware,'' ''pause_video,'' ''load_video,'' and ''play_video.'' In addition, the similarity score between ''problem_check_correct'' and ''stop_video'' increased from approximately 0.994 to 1. This suggests that the PCA process enhances the similarity between these actions, possibly by extracting underlying patterns or by reducing noise in the data. Additionally, even actions with lower initial similarity scores, like ''click_forum'' and ''click_progress,'' show increased similarity after PCA. Moreover, The similarity matrix produced from PCA-based representations of KDDCUP actions in Figure 7 shows increasing clusters of actions with distinctly high similarity scores, like ''navigate'' and ''wiki,'' suggesting their shared characteristics. In essence, this matrix serves as a valuable tool for understanding user behavior patterns, aiding the development of recommendation systems and clustering algorithms.

2) PHASE 2: DOWNSTREAM TASK

The downstream task is the process by which learned representations are applied to downstream tasks for target outcome prediction tasks. To assess the effectiveness of our methodology, we conducted a comparative analysis involving our proposed framework alongside PCA and our framework against a baseline approach utilizing one-hot encoding (refer to Section IV-C1). Specifically, we utilize the SVM model to predict student dropout early on leveraging the latent representations derived from the initial 100 clickstream data. Through this evaluation, we were able to compare the performance of our framework with other methods and evaluate its effectiveness in dropout prediction tasks.

D. RESULTS AND DISCUSSION

This study explored different approaches to predicting student dropout using clickstream data from online learning platforms. We conducted a comprehensive analysis to compare the efficacy of these methodologies, by leveraging different feature extraction and dimensionality reduction techniques. Our findings clarify the effectiveness of these approaches and provide insights into how they may affect predictive modeling in educational settings.

1) BASELINE MODEL WITH SVM

We implemented a baseline model utilizing one-hot encoding for the feature representation. In the baseline model, features were extracted using one-hot encoding, a technique that

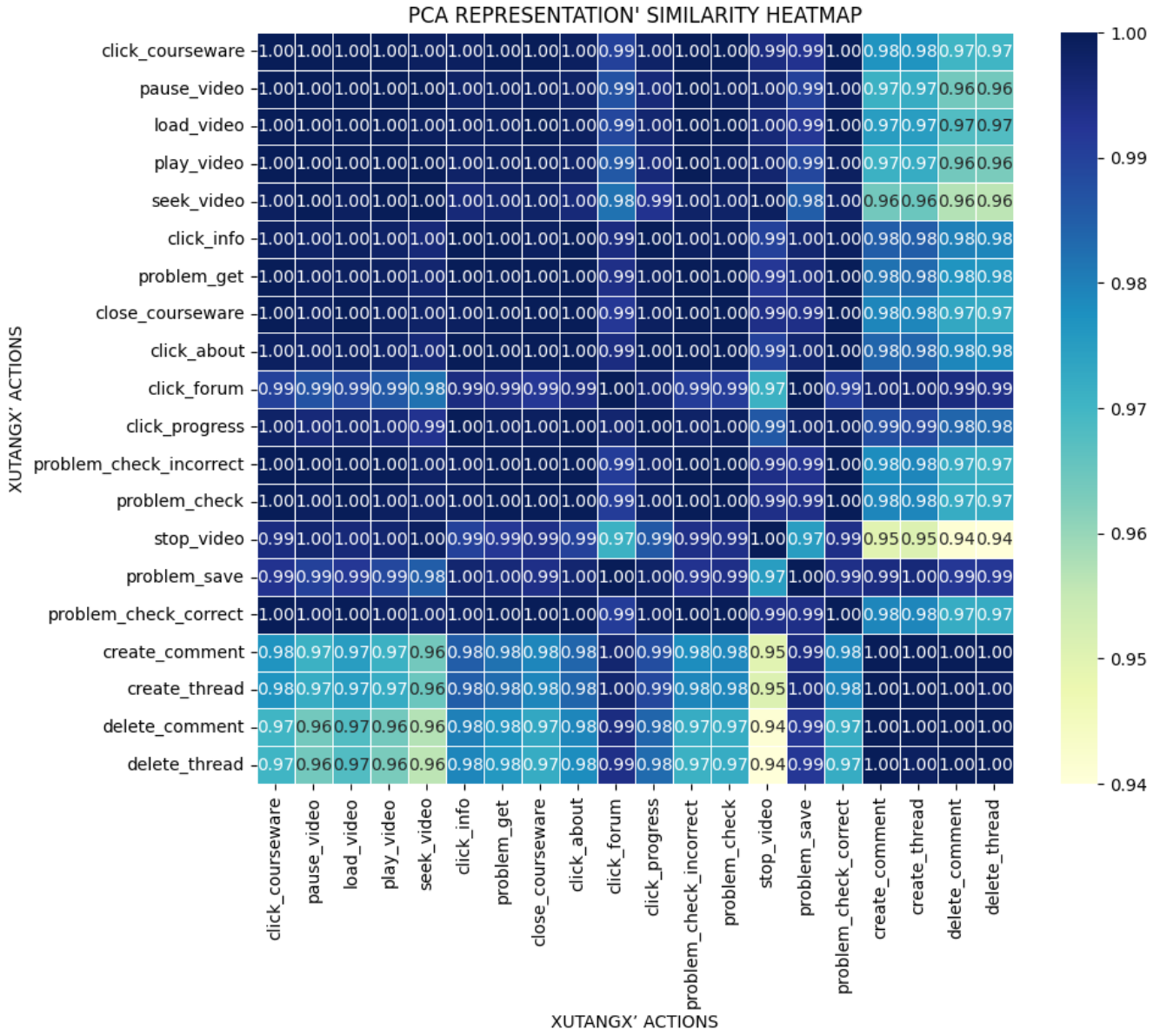

FIGURE 6. Similarity heatmap between XutangX' actions using PCA representation.

represents each event in the clickstream as a binary vector. This process involves iterating through each student's clickstream events, converting them into one-hot encoded vectors, and subsequently aggregating them to form the feature vectors. Padding is applied to ensure uniformity in the vector length, which is set to 100 in this instance. The resultant feature vectors are then organized into a data frame along with their corresponding ground-truth labels. Additionally, the baseline model uses random under-sampling to resample training data to address class imbalance. This process produces resampled data that can be used for further validation. After applying under-sampling to the training data, the class distribution became balanced, with both classes having 6,086 instances each. However, the test data remained unchanged, with 2,580 instances for Class 0 and 8,169 instances for Class 1. The validation step involves employing SVM classification in the resampled training data. The classifier performance is evaluated on the resampled testing data, yielding metrics such as accuracy, F1-score, precision, and recall. The one-hot encoding method in Table 4 was used to assess the performance scores of the baseline model. Despite its simplicity, this approach yielded a moderate predictive performance, with an accuracy of approximately 63.50%, an F1-score of 74.64%, and a recall of approximately 70.23% for the Xutangx dataset. Furthermore, when applied to the KDDCUP dataset, the baseline model yielded an accuracy of 39.08%, F1-score of 38.79%, and recall of 24.27%. Although one-hot encoding provides a simple representation of student

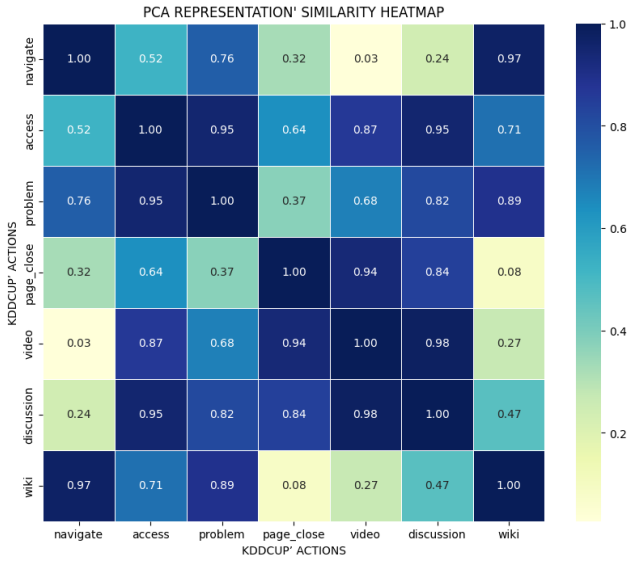

FIGURE 7. Similarity heatmap between KDDCUP' actions using PCA representation.

interactions, it can not capture semantic relationships and leads to information loss. This emphasizes how crucial it is to investigate advanced feature extraction and dimensionality reduction methods, such as PCA representations and skip-gram, to further enhance predictive accuracy, especially in situations involving complex and high-dimensional data, such as clickstream logs.

2) SKIP-GRAM EMBEDDINGS WITH SVM

Our methodology involves utilizing skip-gram embeddings to represent clickstream data, followed by SVM classification. Initially, skip-gram embeddings are employed to extract features from the clickstream data, with the extracted skip-gram feature function iterating through each student's clickstream logs to retrieve word vectors and construct feature vectors. Following this, Random Under-Sampling was utilized to balance the highly imbalanced class distribution observed in the training data. The resampled datasets were then used to train an SVM classifier. The SVM classifier for binary class classification was specified using hyperparameters to optimize its performance. Parameter `class_weight='balanced'` automatically adjusts the weights of the classes based on their frequencies, thereby mitigating the effects of class imbalances. In addition, `gamma='scale'` adapts the kernel coefficient by considering the number of features and their variance. The shape of the decision function is set to `'ovo'` (one-versus-one) to handle binary classification tasks efficiently. The regularization parameter `C=1.0` controls the trade-off between achieving a low training error and low testing error, determining the margin softness. These hyperparameters collectively ensure the SVM classifier is well-tuned to address the class imbalance and achieve powerful performance in binary classification scenarios.

The skip-gram method in Table 4 presents the performance scores of our initial methodology. The reported results indicate promising performance, with an accuracy of approximately 71.88%, an F1-score of 80.03%, and a recall of 74.17%. Additionally, when applied to the KDDCUP dataset, the methodology yielded an accuracy of 80.75%, an F1-score of 87.22%, and recall of approximately 82.83%. The use of skip-gram embeddings enabled the capture of meaningful semantic representations of student interactions, facilitating the identification of patterns indicative of dropout behavior. These findings suggest that the SVM classifier, trained on skip-gram embeddings of clickstream data post-resampling, can effectively identify patterns associated with student dropout.

3) PCA ON SKIP-GRAM REPRESENTATIONS WITH SVM

To advance the predictive modeling pipeline, we integrated PCA into the Skip-gram clickstream representation to strengthen the prediction of student dropout. PCA, a dimensionality reduction technique, was applied to the skip-gram click-stream representation obtained in the previous steps. The number of components for reduction is set to two. The PCA object is trained on the clickstream data to learn transformation parameters, which are subsequently applied to both the training and testing data to yield reduced-dimensional representations. Furthermore, to address the class imbalance, Random Under-Sampling is utilized on the PCA-transformed feature space, generating a resampled training set for validation purposes. SVM classification is employed once again to validate the PCA representation, train the classifier on the resampled training data using PCA-transformed features and evaluate performance metrics on the resampled testing data. By reducing the dimensionality of the skip-gram representations, we aimed to improve the computational efficiency and potentially uncover latent structures in the data. The PCA method in Table 4 shows the performance scores of the proposed framework. Our results confirmed the effectiveness of this approach, with an accuracy of 72.70%, an F1-score of 81.03%, and improvements in the recall metrics. Additionally, when applied to the KDDCUP dataset, the methodology yielded an accuracy of 80.91%, F1-score of 87.42%, and recall of 83.46%. In summary, integrating PCA into a predictive modeling pipeline enhances the identification of student dropout patterns in the clickstream data. PCA provided a compact representation of the clickstream data while preserving its discriminatory power, contributing to improved model performance.

Our comparative analysis highlights the importance of employing advanced feature extraction and dimensionality reduction techniques for predictive modeling in educational settings. By leveraging skip-gram embeddings and PCA, we were able to extract more informative representations of clickstream data, resulting in superior predictive accuracy compared with traditional methods such as one-hot encoding. These findings underscore the value of incorporating machine

learning approaches into educational research and practice, offering opportunities to identify at-risk students early and provide timely interventions to support their learning journeys.

E. QUALITATIVE EXPERIMENTS

We investigate the interpretability of the latent behavioral vectors generated by our framework, focusing on student-level latent vectors extracted from the first 100 behavioral clickstream data. Initially, we fed a K-means clustering algorithm with skip-gram representations of each user's latent vectors from the resampled training data. We computed the silhouette coefficient for different numbers of clusters to identify the optimal clustering configuration based on the Silhouette Analysis. This analysis revealed that the optimal number of clusters is three, with a silhouette score of 0.71, indicating a reasonably well-defined cluster structure. In addition, we utilize 'GloVe [40]: Global Vectors for Word Representation' to represent each event type. We input the GloVe representations of latent vectors of all users into a K-means clustering algorithm. The optimal number of clusters was found to be 3, with a silhouette score of 0.73, suggesting a slightly better-defined cluster structure compared to our framework representations. Given these findings, we chose to further analyze the clustering of GloVe representations owing to the higher silhouette score. We used Plotly's scatter plot to visualize the vectors in a 2D plane and investigate the characteristic behavioral patterns within the visible clusters. Figure 8 illustrates the clusters, providing insights into the distinct behavioral patterns among different student groups. Table 5 presents the detailed clustering results. Analyzing student behavior patterns and course outcomes using three distinct clusters.

- Cluster 0 - Active Forum Users: The majority of these students are characterized by primarily engaging in forum activities. These students showed the highest levels of forum actions but had the lowest engagement in clicking on course information, opening assignments, and loading videos. Observations indicate that this group had a high dropout rate and generally low engagement in overall course activities. It is critical to motivate these students to engage more with the course material to address the high dropout rate within this cluster. This could be accomplished by encouraging the watching of educational videos, doing homework, and increasing interactions with course materials. By encouraging more comprehensive engagement with the course, these clusters of students may become more committed to their learning and decrease the risk of dropping out.
- Cluster 1 - Most video-watching group: This group consisted of students who spent the longest time watching videos, demonstrating a high commitment to studying and understanding the material. These students have a low dropout rate and are serious learners, likely motivated by personal growth through the video content. To support their learning, this could involve additional homework. By catering to their preferred learning styles, these students can continue to grow and succeed in their studies.
- Cluster 2 - Hard Workers: This cluster is identified as the group with the highest overall engagement, particularly in problem-solving activities and interactive course materials. This cluster boasts the lowest dropout rates, with high activity levels correlating with low dropout rates.

Clustering analysis effectively captured meaningful behavioral patterns among students, highlighting the distinct characteristics and engagement levels within each cluster.

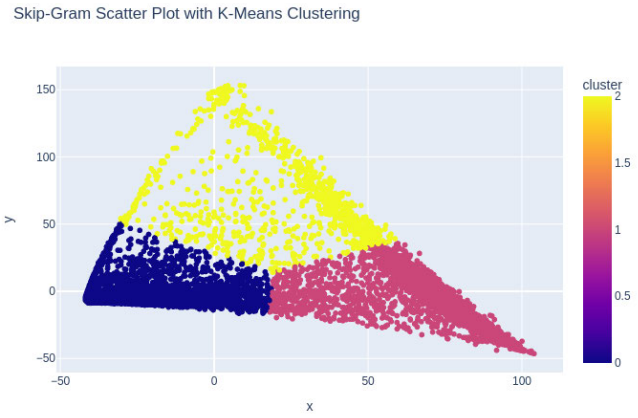

FIGURE 8. 2D visualization of student-level latent vectors using GloVe representations, clustered into three groups.

TABLE 5. Results of the clustering analysis (average).

User Action Categories	Cluster 0	Cluster 1	Cluster 2
Click Actions	7.84	7.99	8.15
Click-about	3.69	3.67	3.56
Click-courseware	17.63	18.34	19.19
Click-forum	4.57	3.93	5.98
Click-info	4.30	4.27	4.21
Click-progress	3.79	3.94	3.87
Close Actions	7.72	7.91	8.79
Close-courseware	7.72	7.91	8.79
Forum Actions	2.46	1.93	1.95
Create-comment	3.17	2.25	2.43
Create-thread	1.45	1.37	1.47
Delete-comment	1.17	1.17	
Delete-thread	1.25	1.00	1.00
Problem Actions	7.63	7.62	8.96
Problem-check	9.40	9.36	13.21
Problem-check-correct	6.73	6.64	8.19
Problem-check-incorrect	5.53	5.30	6.75
Problem-get	9.77	9.93	10.58
Problem-save	4.02	3.98	4.15
Reset-problem	5.83	7.13	2.00
Video Actions	13.27	14.61	14.27
Load-video	6.68	6.86	7.04
Pause-video	15.83	17.14	17.30
Play-video	16.95	18.67	19.42
Seek-video	16.60	19.43	18.00
Stop-video	14.26	16.32	13.04
Total Students	7875	3387	910
Dropout Percentage	62.63%	27.90%	22.97%

Student engagement through clicks and problem-solving attempts were positively associated with passing rates. Strategies that promote active participation, such as interactive content and problem-solving activities, are crucial for reducing dropout rates and improving overall success in online courses. Therefore, creating an engaging and interactive learning environment is key to enhancing student outcomes and ensuring success in MOOCs. The results underscore the effectiveness of our approach in identifying and understanding student behavior, providing actionable insights to support their academic journeys.

V. CONCLUSION

In this study, we explored the application of natural language processing techniques for clickstream data representation to improve the prediction of student dropout. We investigated the effectiveness of using the skip-gram model in machine learning. The baseline approach utilizing one-hot encoding provided a foundational understanding of student interactions but fell short of capturing semantic relationships and dealing with high dimensionality. Despite its moderate predictive performance, its sensitivity to dataset characteristics was evident when tested on the KDDCUP dataset, thereby emphasizing the need for more powerful methods. In contrast, leveraging skip-gram embeddings with SVM classification yielded promising results. This approach facilitates the identification of dropout patterns by capturing meaningful semantic representations of student interactions. Integrating PCA into the skip-gram representation pipeline provides additional benefits, enhancing computational efficiency. This resulted in an improved predictive accuracy, confirming the effectiveness of dimensionality reduction techniques in dropout prediction. The results indicate that the methodology incorporating skip-gram embeddings with PCA achieved the highest predictive accuracy. Specifically, on the Xutangx dataset, the approach yielded an accuracy of approximately 72.70%, and an F1-score of approximately 81.03%. This suggests that skip-gram embeddings combined with PCA for dimensionality reduction, can effectively capture meaningful patterns in clickstream data related to student dropout prediction. Additionally, when applied to the KDDCUP dataset, the methodology exhibited an even higher performance, with an accuracy of 80.91% and an F1-score of 87.42%. These findings underscore the robustness and generalizability of the proposed framework across diverse datasets, highlighting its potential for predicting dropout in educational settings.

Our research findings demonstrate the significant impact of natural language processing techniques on student performance prediction in MOOCs. By extracting more informative representations of clickstream data, our approach enables deeper comprehension of student behavioral patterns, thus facilitating early intervention strategies. In future work, incorporating deep learning models could further enhance the predictive capabilities of the proposed approach. Exploring deep learning architectures tailored to handle sequential data, such as RNNs or LSTMs, could offer valuable insights and potentially outperform traditional machine learning algorithms.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number IFPRC-039-126-2020 and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

REFERENCES

[1] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C. Ding, S. Wei, and G. Hu, "Exercise-enhanced sequential modeling for Student performance prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2018, no. 1, pp. 1–22.

[2] Y. K. Salal, S. M. Abdullaev, and M. Kumar, "Educational data mining: Student performance prediction in academic," *J. Eng. Adv. Tech.*, vol. 8, no. 4C, pp. 54–59, 2019.

[3] J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue, "Modeling and predicting learning behavior in MOOCs," in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, Feb. 2016, pp. 93–102.

[4] W. Bao, "COVID-19 and online teaching in higher education: A case study of Peking university," *Hum. Behav. Emerg. Technol.*, vol. 2, no. 2, pp. 113–115, Apr. 2020.

[5] C. Impey and M. Formanek, "MOOCS and 100 days of COVID: Enrollment surges in massive open online astronomy classes during the coronavirus pandemic," *Social Sci. Humanities Open*, vol. 4, no. 1, 2021, Art. no. 100177.

[6] G. Korkmaz and Ç. Toraman, "Are we ready for the post-COVID-19 educational practice? An investigation into what educators think as to online learning," *Int. J. Technol. Educ. Sci.*, vol. 4, no. 4, pp. 293–309, Sep. 2020.

[7] D. Liu, Y. Zhang, J. Zhang, Q. Li, C. Zhang, and Y. Yin, "Multiple features fusion attention mechanism enhanced deep knowledge tracing for Student performance prediction," *IEEE Access*, vol. 8, pp. 194894–194903, 2020.

[8] A. Ramesh, D. Goldwasser, B. Huang, H. Daume, and L. Getoor, "Learning latent engagement patterns of Students in online courses," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1–26.

[9] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods," in *Proc. EMNLP Workshop Anal. Large Scale Social Interact. MOOCs*, 2014, pp. 60–65.

[10] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *J. Mach. Learn. Res.*, vol. 15, no. 57, pp. 1959–2008, 2014.

[11] M. Berland, R. S. Baker, and P. Blikstein, "Educational data mining and learning analytics: Applications to constructionist research," *Technol., Knowl. Learn.*, vol. 19, nos. 1–2, pp. 205–220, Jul. 2014.

[12] D. M. West, "Big data for education: Data mining, data analytics, and web dashboards," *Governance Stud. Brookings*, vol. 4, no. 1, pp. 1–10, 2012.

[13] M. Ding, K. Yang, D.-Y. Yeung, and T.-C. Pong, "Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses," in *Proc. 9th Int. Conf. Learn. Anal. Knowl.*, Mar. 2019, pp. 135–144.

[14] T.-Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, "Behavior-based grade prediction for MOOCs via time series neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 5, pp. 716–728, Aug. 2017.

[15] L. Yang, J. Chen, Z. Wang, W. Wang, J. Jiang, X. Dong, and W. Zhang, "Semi-supervised log-based anomaly detection via probabilistic label estimation," in *Proc. IEEE/ACM 43rd Int. Conf. Softw. Eng. (ICSE)*, May 2021, pp. 1448–1460.

[16] Md. H. Rahman and Md. R. Islam, "Predict Student's academic performance and evaluate the impact of different attributes on the performance using data mining techniques," in *Proc. 2nd Int. Conf. Electr. Electron. Eng. (ICEEE)*, Dec. 2017, pp. 1–4.

[17] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk Students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[18] C.-H. Yu, J. Wu, and A.-C. Liu, "Predicting learning outcomes with MOOC clickstreams," *Educ. Sci.*, vol. 9, no. 2, p. 104, May 2019.

[19] S. Qu, K. Li, B. Wu, S. Zhang, and Y. Wang, "Predicting Student achievement based on temporal learning behavior in MOOCs," *Appl. Sci.*, vol. 9, no. 24, p. 5539, Dec. 2019.
[20] M. Ding, Y. Wang, E. Hemberg, and U.-M. O'Reilly, "Transfer learning using representation learning in massive open online courses," in *Proc. 9th Int. Conf. Learn. Anal. Knowl.*, Mar. 2019, pp. 145–154.
[21] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, "MOOC dropout prediction using FWTS-CNN model based on fused feature weighting and time series," *IEEE Access*, vol. 8, pp. 225324–225335, 2020.
[22] Y. Xie, "Student performance prediction via attention-based multi-layer long-short term memory," *J. Comput. Commun.*, vol. 9, no. 8, pp. 61–79, 2021.
[23] Y.-W. Chu, E. Tenorio, L. Cruz, K. Douglas, A. S. Lan, and C. G. Brinton, "Click-based Student performance prediction: A clustering guided meta-learning approach," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2021, pp. 1389–1398.
[24] A. Scarlatos, C. Brinton, and A. Lan, "Process-BERT: A framework for representation learning on educational process data," 2022, *arXiv:2204.13607*.
[25] F. Han, "Representation learning on unstructured data," Ph.D. dissertation, Dept. Comput. Sci., Univ. California, Santa Barbara, CA, USA, 2016.
[26] A. S. Imran, F. Dalipi, and Z. Kastrati, "Predicting Student dropout in a MOOC: An evaluation of a deep neural network model," in *Proc. 5th Int. Conf. Comput. Artif. Intell.*, Apr. 2019, pp. 190–195.
[27] F. Xiong, K. Zou, Z. Liu, and H. Wang, "Predicting learning status in MOOCs using LSTM," in *Proc. ACM Turing Celebration Conf.*, May 2019, pp. 1–5.
[28] S. Yin, L. Lei, H. Wang, and W. Chen, "Power of attention in MOOC dropout prediction," *IEEE Access*, vol. 8, pp. 202993–203002, 2020.
[29] L. Qiu, Y. Liu, Q. Hu, and Y. Liu, "Student dropout prediction in massive open online courses by convolutional neural networks," *Soft Comput.*, vol. 23, no. 20, pp. 10287–10301, Oct. 2019.
[30] W. Wang, H. Yu, and C. Miao, "Deep model for dropout prediction in MOOCs," in *Proc. 2nd Int. Conf. Crowd Sci. Eng.*, Jul. 2017, pp. 26–32.
[31] A. A. Mubarak, H. Cao, and I. M. Hezam, "Deep analytic model for Student dropout prediction in massive open online courses," *Comput. Elect. Eng.*, vol. 93, Jul. 2021, Art. no. 107271.
[32] N. Wu, L. Zhang, Y. Gao, M. Zhang, X. Sun, and J. Feng, "CLMS-Net: Dropout prediction in MOOCs with deep learning," in *Proc. ACM Turing Celebration Conf.*, May 2019, pp. 1–6.
[33] Y. Zhang, L. Chang, and T. Liu, "MOOCs dropout prediction based on hybrid deep neural network," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2020, pp. 197–203.
[34] T. Anuradha, A. Tigadi, M. Ravikumar, P. Nalajala, S. Hemavathi, and M. Dash, "Feature extraction and representation learning via deep neural network," in *Computer Networks, Big Data and IoT*. Cham, Switzerland: Springer, 2022, pp. 551–564.
[35] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2013, pp. 746–751.
[36] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometric Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[37] Q. Fu, Z. Gao, J. Zhou, and Y. Zheng, "CLSA: A novel deep learning model for MOOC dropout prediction," *Comput. Electr. Eng.*, vol. 94, Sep. 2021, Art. no. 107315.
[38] D. Sun, Y. Mao, J. Du, P. Xu, Q. Zheng, and H. Sun, "Deep learning for dropout prediction in MOOCs," in *Proc. 8th Int. Conf. Educ. Innov. Through Technol. (EITT)*, Oct. 2019, pp. 87–90.
[39] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in MOOCs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 517–524.
[40] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2014, pp. 1532–1543.

SHROOQ AL AMOUDI received the bachelor's degree in computer science from the University of Umm Al-Qura (UQU), Makkah, Saudi Arabia. She is currently pursuing the master's degree in computer science with King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include machine learning, deep learning, and artificial intelligence.

AREEJ ALHOTHALI received the Ph.D. degree in computer science, specializing in artificial intelligence from the University of Waterloo, Canada, in 2017. She is currently an Associate Professor with the Faculty of Computer Science and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include machine learning, deep learning, natural language processing, computer vision, intelligent agent systems, and affective computing.

RSHA MIRZA received the Ph.D. degree in computer science, specializing in software engineering from the University of Colorado Boulder, USA, in 2019. She is currently an Assistant Professor with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include big data, social media analysis, software engineering, groupware (collaborative systems) and CSCW, HCI, data science, data mining and machine learning, and artificial intelligence.

HUSSEIN ASSALAHI received the Ph.D. degree in TESOL from the University of Exeter, U.K., in 2016. He is an Associate Professor of TESOL with the English Language Institute, King Abdulaziz University. His research interests include second language teacher education, language teacher professional development, professionalism, second language acquisition, and critical pedagogy.

TAHANI ALDOSEMANI is a Professor of educational technology. She is the Director of cultural higher education. Her previous roles include a Professor of educational technology with Prince Sattam Bin Abdulaziz University, a member of the University's Council, and the Vice Dean of information technology and distance education. She was an Advisor of the Minister of Education, focusing on e-learning and international cooperation. She also served as the Co-Chair for the G20 2020 Education Group. She has received several international awards and recognitions in educational research and has many publications in educational technology and digital transformation in education. She has implemented many successful initiatives in education presented at conferences, seminars, and workshops.