IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 19, 2022, accepted May 1, 2022, date of publication May 3, 2022, date of current version May 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172438

Utilizing Social Clustering-Based Regression Model for Predicting Student's GPA

YOMNA M. I. HASSAN[1], **ABEER ELKORANY**[2], **AND KHALED WASSIF**[2]
[1]Faculty of Computer Sciences, Misr International University, Cairo, Egypt
[2]Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt

Corresponding author: Yomna M. I. Hassan (yomna.ibrahim@miuegypt.edu.eg)

ABSTRACT The importance of e-learning has exceeded expectations over the past decade. Accordingly, several systems have been developed in completing intelligent assistive tools where students' behavior can be tracked and followed with suitable recommendations to enhance students' performance. This paper has two main objectives. First, the Community of Inquiry framework (CoI) is utilized as one of the most prominent student behavioral modeling to select features that best represent the students. According to experts' annotation, this study filters students' measured attributes from the StudentLife dataset to the CoI model, focusing on social presence. Second, the research looks at improving the accuracy and runtime of the Grade Point Average (GPA) prediction by introducing a hybrid model that combines combining k-means clustering phase based on student similarity with regression-based prediction. The clustering was performed on both static and Spatio-temporal (spatial time -series) students' attributes. Results show that LassoCV outperforms other regression techniques such as Standard Linear, Lasso, and Ridge Regression with an RMSE averaged around 0.15 and an average Adjusted R2 of 0.935 overall trials. Selecting the features according to the CoI reduces the number of features by 62.8%. Time-series clustering on its own was not beneficial; however, when conducted with the selection phase, it raised the quality of the model achieved by 2-3%.

INDEX TERMS Time-series clustering, k-means, community of inquiry, GPA prediction, DTW similarity, student modeling.

I. INTRODUCTION

With the increase of online learning due to recent pandemic conditions comes the need to effectively analyze student data to enhance the learning experience, more specifically with the existence of much student data that can be measured through various platforms [1], [2]. This increase in data availability led to more intelligent educational systems, especially with the move to online learning. Intelligent education systems [3], [4] have been an assistive measure for online and blended education. These intelligent systems are defined as spaces where technology and environmental factors are taken into account to improve students' academic performance [3]. Intelligent education systems involve multiple layers, including but not limited to as it is a repeatedly evolving field:

- Students' profile and behavioral analytics [5].
- Feedback Cycle and recommendation based on analytics [6].
- Content creation and personalization [7].
- Futuristic performance prediction [8].

As gathered from previous studies [9], [10], most of the conducted research relies heavily on being able to model and represent the student effectively. It also became more integral to utilize existing time-series based logs and trails to model the student [11]–[13] instead of relying on constant survey collection that specifically targets student modeling, specifically students' perceptions and behavior [14]–[16].

Researchers increasingly depend on smartphones and wearables to continually monitor students' everyday lives. This monitoring aims to detect numerous elements that can affect the students. These elements include but are not limited to phone call habits connected with their health and well-being and academic success [17].

According to the findings, various factors, including academic and non-academic characteristics, influence student academic progress [18]. Earlier publications focused on utilizing automatically gathered data from multiple sources around the student, such as Mobile sensors, wifi checkpoints,

The associate editor coordinating the review of this manuscript and approving it for publication was Ehab Elsayed Elattar.

Class QR check codes, and surveys. These attributes are generically gathered for student health purposes to try and predict the students' performance [19], [20]. However, previous work utilized most of the data available about the student without a pre-planned structured selection of the most influencing data. Furthermore, While numerous studies attempted to develop intelligent classifiers for anticipating student achievement, they overlooked the importance of identifying the key factors that lead to the achieved performance [8]. It also must be considered that capturing part of the student data instead of utilizing all the data gathered can be ineffective. Any model of learning can be considered wrong in specific ways because it cannot capture the full complexity of the actual student's mind [21]. Studying various student modeling techniques for generalization purposes over various learning platforms and datasets became critical to understand which model is the most effective in capturing the student's data and surroundings effectively.

The research in this paper considers the definition of student modeling as the learner representation, involving both the characteristics selected as the representation and the techniques used to utilize these characteristics to enhance the learner's learning environment, as described in [22]. More details are presented in the literature review.

The characteristics defined can be regrouped from various data sources and compiled into [21], [23]:

- Learner profiles: i.e., visual representations of their actions.
- Mastery of knowledge on a specific domain or topic.
- Cognitive and meta-cognitive abilities, as well as affective moods.

Many studies focused on how various behavioral, psychometric, or educational models may influence students' learning process, mainly how these models can be utilized in predicting futuristic outcomes of the student [23], [24], as test scores are used as a proxy for learner proficiency. The paper evaluates the performance of the student modeling performed by analyzing the prediction outcome on a publicly available dataset.

This study is considered different from other smaller scaler studies as it considers a more generalizable student representation that can quickly be adopted by more than one dataset. In addition, a comparison is conducted to present the improvement achieved by combining multiple techniques such as time-series clustering and classification to enhance the prediction accuracy and minimize the number of features required from the dataset. This comparison can be achieved by studying how including students' social aspects within the student model can enhance students' outcome prediction. These social aspects are considered within and outside academic boundaries and are collected through automatically measured student trials.

The study also shows how specific characteristics of the students deemed necessary are selected according to the existing educational framework, in particular Community of Inquiry, can contribute to the outcome prediction and minimize the required dataset. The model's effectiveness is studied in the realm of predicting student cumulative academic performance overall semesters; futuristic work may reveal more relations regarding the performance over a single semester. In addition, student similarity based on time-series clustering will be analyzed as a pre-prediction process for enhancing prediction performance and minimizing utilized features and attributes.

In summary, the main objectives of this paper are:

- Validate previous work conducted on a publically available dataset and utilize the best performing regression model as a baseline for comparison of other regression models, explicitly using the complete attributes presented from the dataset.
- Select student's attributes to the Community of inquiry framework to model and represent the student with the most critical indicators while ignoring unnecessary factors.
- Utilize time series clustering on time series attributes representing social presence in the Community of Inquiry framework and performing time-series similarity-based techniques to form clusters of students. This assists in finding similar students concerning behavior and how these clusters may enhance predicting student outcomes in a hybrid 'cluster-then-predict' methodology.

The structure of this paper is divided into four main sections, starting with section II (Literature review) describing related research conducted in student modeling, Community of inquiry, and GPA prediction. Section III includes the methodology and the detailed description of techniques applied and how the comparative analysis will be conducted and evaluated. Eventually, the results are presented and discussed in section V to ensure the validity of the methodology represented. Potential future paths are detailed within the conclusion section to pave the way for other researchers.

II. LITERATURE REVIEW

This section includes a detailed history of the research conducted in the fields correlated to the objectives achieved by the paper. The first sub-section covers background about student modeling, and various characteristics and techniques are considered to model the student effectively. The following sub-section includes details about the Community of Inquiry (CoI) framework, which selects relevant attributes and minimizes the students' characteristics. Finally, a sub-section includes details of the history of GPA prediction, including how clustering was used in this field before.

A. STUDENT MODELLING

Student modeling is a modeling technique that measures the current knowledge state of the student [25]. According to previously conducted research [22], the learner/student modeling process is divided into two main partitions:

- The characteristics of the students to be studied
- The technique/process used to model the student to achieve a better learning environment.

TABLE 1. Learner's Characteristics as appeared in literature.

Name	Definition
Learner Profile	The learner profile comprises uninterpreted static data, such as age, gender, and name.
Knowledge	Level of knowledge, competencies, skills, blunders, misunderstandings, and forgetting
Cognitive Characteristics	Learning styles, working memory capacity, thinking styles, cognitive states, and learner behavior are all factors to consider.
Social Characteristics	Interactions with others, culture, social style (collaborative vs. lonely), and available time are all considered factors.
Motivation	Interests, learning goals, engagement and affect.

Table 1 identifies Learner characteristics discussed according to the review [22]. The review paper showed various characteristics studied in previous research, including the learner's static profile, knowledge, cognitive ability, social characteristics, and motivation. The review describes these characteristics with details to entail all possible descriptions in various publications. These characteristics are represented through multimodal data, which refers to interaction traces that occur across multiple communication channels [26]. Multimodal data capture aspects of learning that are difficult to observe with the naked eye or through self-reported data (e.g., mental exertion, emotional states).

Some applications and techniques applied to these characteristics include clustering/classification/prediction, overlay modeling, and ontology-based learner models. These characteristics have been used separately and in a hybrid form to enhance the feedback cycle to the student and eventually create an automated satisfactory learning environment.

The characteristics represented above have been presented to affect various datasets' outcomes significantly. However, the required outcome is to select student features to a generic model representing the student. Models have been previously utilized or explicitly designed to select automatically measurable characteristics to represent the student [27]. However, generalized models that encompass more than one type of characteristics have not been used in this context.

For this purpose, various general representations of the students used in other contexts have been explored. One of the most widely used representations in online education is the Community of Inquiry model [28]. More details about this model and how it is used in the context of the educational system are present in the following sub-section.

B. COMMUNITY OF INQUIRY

In online learning, presence is seen as a critical notion. Presence is identified people's ability to project their qualities/behaviors and thoughts into the community. People presenting themselves to other participants as real people is defined as ''presence'' in the online course.

The Community of Inquiry (CoI) framework, a widely referenced model that describes three interrelated elements of presence: social, teaching, and cognitive, is one technique to investigate 'presence' within online/blended learning communities (Figure 1). When all three aspects come together, students can have a profound and meaningful learning experience [29].

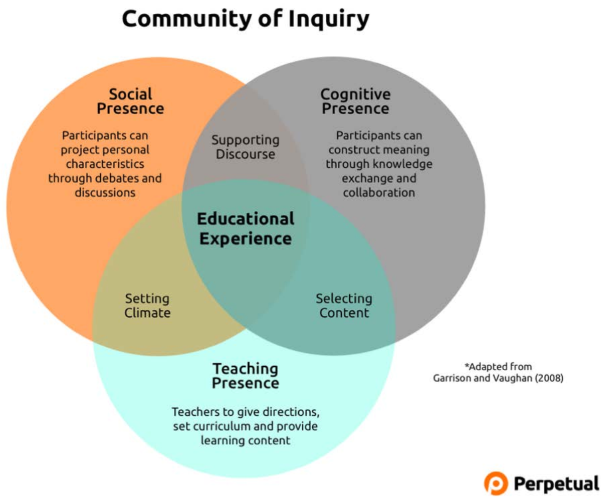

Community of Inquiry
Social Presence
Participants can project personal characteristics through debates and discussions
Cognitive Presence
Participants can construct meaning through knowledge exchange and collaboration
Supporting Discourse
Educational Experience
Setting Climate
Selecting Content
Teaching Presence
Teachers to give directions, set curriculum and provide learning content
*Adapted from Garrison and Vaughan (2008)
Perpetual

FIGURE 1. Community of Inquiry Framework [30].

Concerning educational theories about e-learning that were used in international high-impact scientific journals from 2009 to 2018, CoI emerges as the most relevant theoretical framework in these publications [31].

CoI as a framework has been previously used in different contexts. The main focus of the research related to CoI was:

- How to utilize CoI based surveys in detecting/predicting specific characteristics/presences within students [32]–[34].
- How to utilize CoI indicators to perform either manual/automated tagging of content related to students, such as transcripts/messages, to validate the existence of certain presence in between students [35], [36].

The closest similar system to the design process suggested in the paper in-hand was mentioned as a project proposal [37]. The research suggested in the proposal is based on utilizing an automated approach in detecting CoI models between students. Their suggestion is similar to previous research that focused on analyzing messages and transcripts to detect certain presences [38], [39], such as cognitive presence from how ''wordy'' is a specific phrase. What they added in their proposal is giving the ability for the student to visualize their presence level overtime and measure the influence of this visualization on the students' outcome [37].

The main differences between the work presented in this paper and previously presented work are:

- The automatic detection of CoI is not based on messaging or connections on social networks. It selects features matching the presence based on existing measurable data about the students from sensors or classroom data.
- The selected features are then used to predict student outcomes such as GPA. This model/representation of the students can be used in other areas of the student learning cycle, not just outcome prediction.

The selection process conducted in this research between students' features/attributes and the CoI framework is performed through matching between attribute names, attributes descriptions, and the CoI indicators [40]. CoI indicators have been utilized as tagging/mapping tools (Text-to-Text) Mapping in several research papers. Markers of presences such as teaching/cognitive/social presences are identified in messages/transcripts related to the students to ensure the existence of CoI within the learning environment [35], [41]–[43]. In this paper, instead of mapping transcripts or messages as in previous research, a selection is conducted from the attributes measured automatically.

The indicators have been used in both automatic/ non-automated mapping of content. However, up to our knowledge, the relation between the students' mapping and student outcomes has not been studied comprehensively. The following section will detail how GPA prediction is performed in general and the CoI framework to better understand the field.

C. GPA PREDICTION

The relation between both academic and extracurricular life of the student has been previously discussed [44]. Various approaches are entailed, specifically ones that have been tested previously to enhance the prediction of students' performance. In addition, the difference in the features available, the technique used, and the evaluation criteria will be described.

Up to our knowledge, a single paper was published on the same dataset utilized here and within the scope of work (Student outcome prediction) [45]. Their work utilized multiple features gathered from surveys, mobile sensors, and other geographically placed sensors and calculated the correlations between various features and each other. Later on, the research was expanded to focus on GPA prediction [46]. They have used Pearson's correlation as a feature selection method, which finally represented their model for the student. The mean absolute error of their predicted cumulative GPA is 0.179.

As mentioned, the data collected was a mixture between the student's personal life and their interaction and effort within their course work. Previous work entails as well research that has been conducted on either the course work details separately [47], which achieved maximum accuracy of 80.47%, and [48] with an average mean square error of 0.2 at the end of a quarter. In other cases, research focused only on students' static profile as a GPA predictor [49], even though it was potentially less accurate.

1) OUTCOME PREDICTION THROUGH CoI MODEL

Various outcomes indicators can be collected and predicted to validate the educational systems' success through the learning process.These outcomes include but are not limited to student engagement, mood, and general health, drop-out rate, patterns and similarities, and academic performance. This sub-section mentions research that utilized the CoI concept to predict or classify students according to specific measurements.

The first discussed study examined the effects of the Learning Management System (LMS) on student outcomes [50]. Quantitative analyses of survey data involved a wide range of courses and faculty, examining the effects of LMSs on essential learning outcomes. Results supported critical aspects of the CoI model, indicating the importance of technology in facilitating all three presences of the CoI framework and satisfaction with online courses. Correlation analysis showed that technology made communication easy (Beta = 0.15, $p < .05$), In addition students reading all online elements had an even stronger effect (Beta = 0.30, $p < .001$).

Following the previous study, in 2015, trace data was extracted automatically from Moodle's PostgreSQL database and consisted of almost 200,000 log records of different student activities [51]. The students posted 1747 messages in total, which, together with the LMS trace data – represented the primary data source for this study. Manual tagging of these traces was conducted. The tagging focused only on the cognitive presence aspect of the CoI model. MANOVA analysis was performed. Results revealed six technology-use profiles associated with different levels of cognitive presence (shown in Figure 2). The same researcher repeated the same process over transcripts data later in 2017 [52]. A follow-up on this work was again utilizing Massive Open Online Courses (MOOCs) interaction log data [53]. The Community of Inquiry survey instrument, administered as part of the post-course survey, was used to measure student perceived levels of the three presences. Cluster analysis revealed three students with unique study strategies: limited users, selective users, and broad users.

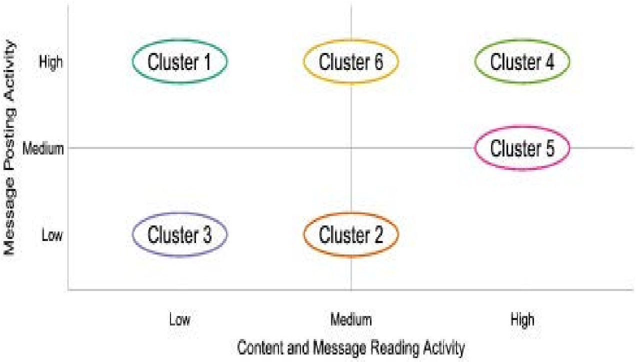


FIGURE 2. **Cluster matrix: activity focus and activity level [51].**

In 2016, the CoI survey was utilized for outcome prediction [54]. Results from path analysis confirmed that only the cognitive element directly correlated with continuous academic-related online performance and satisfaction. Similar work was conducted in 2021 [55].

The relation between online behaviors (i.e., four types of learning behaviors) and learning performance over MOOCs was studied next [56]. This study investigates the community of inquiry model using questionnaires and learning behavior data collected from a learning platform. The correlation and stepwise linear regression analysis identified a correlation between learning experience with CoI learning presences, average correlation results was ($r = .557, p < 0.01$). However, student outcomes were not directly correlated to CoI, average correlation is shown to be ($p = 0.025 < 0.05$).

Regarding clustering, it has been used as a substitution for the prediction process (identifying the student's level of performance instead of identifying the actual value of the outcome). Clustering was used to classify students according to their performance level [57]–[59]. A hybrid model, such as 'cluster-then-predict' as shown in [60], has not been suggested before in the field of student outcome prediction.

D. RESEARCH GAPS

To sum up, the existing literature has examined the relationship between social presence, teaching presence, and cognitive presence. However, the investigations were primarily based on questionnaire data rather than actual behaviors. Even the research considering temporal behavioral has not studied spatio-temporal factors' influence on students' performance. Noticing that research is more direct towards manual annotation than automatic annotation (which is a new research trend, due to support that manual performs better than automated [61]).

Considering the previous review, we notice no concrete analysis of several potential combinations of student characteristics and modeling techniques tested on the same dataset. This paper offers an opportunity to experiment with various student characteristics based on the CoI framework selection, focusing on social presence. It also presents two of the most prominent modeling techniques: clustering, and predictive modeling. Testing this variation on a dataset with a large number of features can explain how various behavioral and technical models can benefit in understanding students' performance, and futuristically assist in the process of enhancing the students' learning environment through feedback.

Another difference in this paper is how the social presence of the student is identified. What this paper identifies as social presence is divided into two parts:

- Location of students over time, measured by the GPS sensor. The similarity in such trends means similar social patterns [62], [63].
- Attributes representing social presence as selected according to the CoI framework.

To summarize, the main contributions of the paper can be detailed as:

- Using the Community of Inquiry framework to achieve a new representation of the student. This representation is achieved by encoding students' attribute names to CoI indicators and accordingly performing feature selection.
- Integrating Time-series clustering based on Spatio-temporal features (representing social presence) with standard regression models, and how this clustering can enhance the regression-based prediction performance.

The following section gives a detailed construction of the methodology used to achieve the objectives defined in the introduction.

III. MATERIAL AND METHODS

A survey was conducted to detail multiple ways a student can be modeled. Out of these techniques, it is viewed that relating measurable data from the student to specific psychology-based models was used frequently. The methodology is divided into three main sections, representing: How characteristics are selected according to the CoI framework, performing predictive analysis of GPA through regression, and utilizing time series data for grouping students with similar characteristics.

In this paper, the selection of students' measured behavior according to pre-existing model was performed through experts measuring the similarities between attribute names/tables names and CoI indicators. This idea evolved through similar techniques performed through tagging messages and transcripts of students to similar indicators. The methodology is conducted through the following steps:

- Step 1: Gathering potential indicators for the CoI model from previous research into a dataset. This dataset includes all paraphrasing possibilities for each of these indicators [40], [64]–[70].
- Step 2: Tagging each attribute to the indicator most suitable. Noting that an attribute can be tagged by more than one indicator, and in case of non-existence of suitable indicator, the attribute is eliminated.
- Step 3: In case there is confusion regarding one of the indicators, refer to the CoI survey description of the presences to confirm the most suitable presence.
- Step 4: Apply the 'cluster-then-predict' concept using time-series clustering before the prediction process and compare the results with the expected prediction results.

After validating the hypothesis regarding each of the presences' involvement in modeling a student behavior for GPA prediction, which was performed by utilizing only the selected features for the GPA prediction process, the involvement of potential time series analysis is studied. Time-series analysis incorporates the factor of detecting similarity patterns over time in students' behavior. This feature of time-series analysis was shown previously to be exceedingly important in detailing the relations between students and their outcomes and how the students may potentially perform in the future. This paper focuses mainly on time-series based attributes related to social presence construct within the CoI model. The attributes studied for this part include mainly mobile-sensor-based collected data. For the analysis, Uber's Kepler.gl [71] is utilized first to view potential distributions

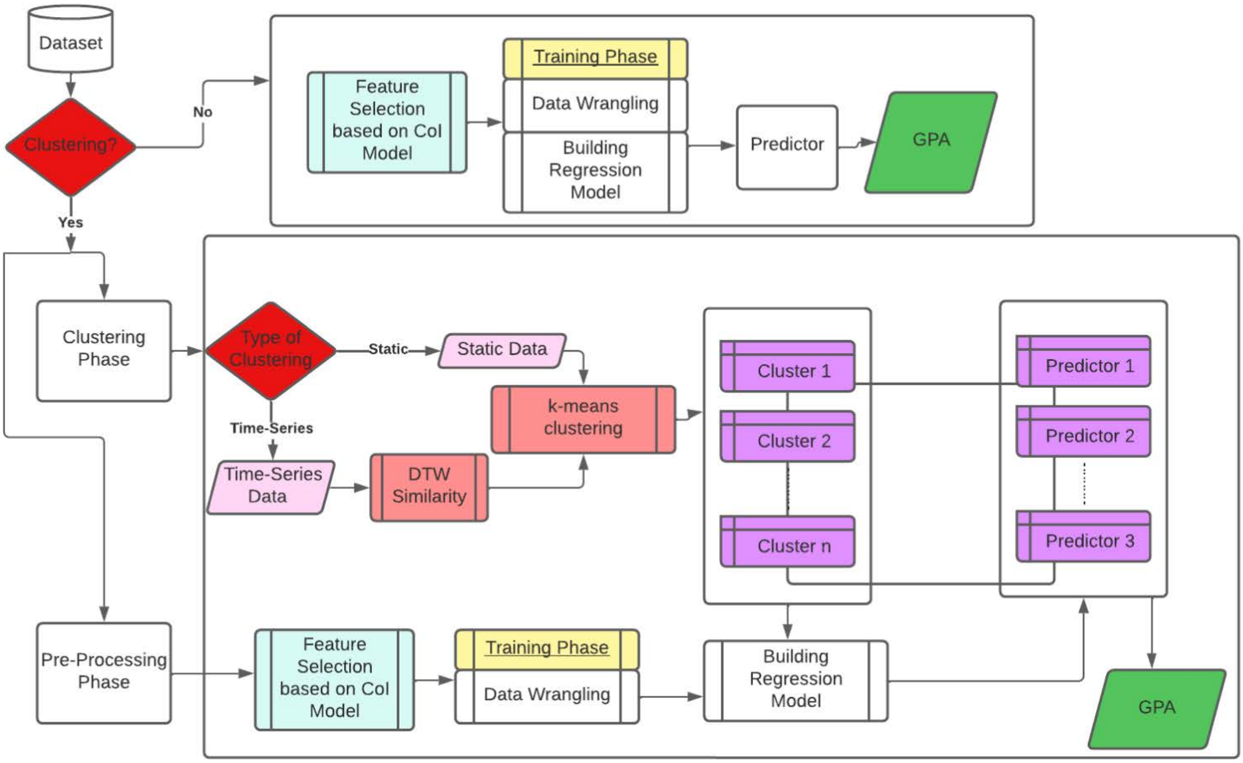


FIGURE 3. System Overview.

of students' behavior over various maps. This assists in identifying preliminary distributions and patterns that can be confirmed through the utilized clustering mechanism. Figure 3 summarizes the methodology. The algorithm structure of the suggested technique is represented in Algorithm 1.

Below, sub-sections including details about the Dataset utilized, the technicalities behind the prediction process, and the 'cluster-then-predict' technique are presented to understand the process further.

A. DATASET

Many of the conducted work is focused on datasets that have been gathered privately in universities/schools. Upon covering available public datasets for educational analytics, several comprehensive datasets were available such as the Open University Learning Analytics Dataset [72], and the LAK dataset [73]. Those publicly available datasets have a limited number of attributes, with no variety in the type of attributes, and with a focus on features involving standard online learning platforms. Most prominently, previous research targets utilizing the whole range of attributes while performing standard feature selection such a Principal Component Analysis (PCA).

For this work, the StudentLife dataset was selected [45]. The suitability of this dataset was due to its rich content. This content involves how the student behaves before, during, and after the learning flow within an academic semester. This dataset would present us with a large selection of features from which the feature selection technique can be evaluated accurately.

The Student Life dataset has been collected to realize the causes of various behaviors from the students. It was collected through analyzing mobile sensory data, standard pre-course and post-course surveys, and academic courses information such as GPA and class timings. The dataset is available for 48 students over ten weeks in the spring 2013 term. In order to get a better perspective about the number of features available per student within this dataset, the statistics within the dataset's Kaggle page were viewed [74]. The statistics of the dataset showed that there are 218 features presented over 1983 tables. Details of the structure of the dataset and how it was collected is shown in Figure 4, which was shown in [75]. Sample of the amount of collected data variation overtime for certain parameters are presented in Figures 5,6.

This dataset has not only been used for GPA prediction or related educational experiments [46]. Others have used the dataset to evaluate mental health progress, or mood variations [76]–[79] due to the variety in the gathered attributes.

B. PRE-PREDICTION PROCESSING USING STATIC AND TIME-SERIES CLUSTERING

In this paper, various pre-processing steps are applied to enhance the data. Initially, values in JavaScript Object Notation (JSON) files of the dataset were aggregated in

```
Algorithm 1 Compute GPA
  Input: Students Static Attributes $Sxs$, Students Time-Series Attributes $Sxt$;
  Apply /Data Wrangling over student data
  if Clustering then
      if Time Series Clustering then
          Compute Similarity of timeseries based attribute between students using DTW
          IdentifyClusters based on similarity using kmeans algorithm
          Calculate Number of Clusters $Ns$: Optimum Silhouette Score
      else
          Identify Clusters based on static attribute using kmeans algorithm
          Calculate Number of Clusters $Ns$: Optimum Silhouette Score
      end if
      For each cluster in the $Ns$ Clusters
      Apply Selected Regression models over students within cluster: $GPA$
  else
      Apply Selected regression models over all students at once: $GPA$
  end if
  return $GPA$
```

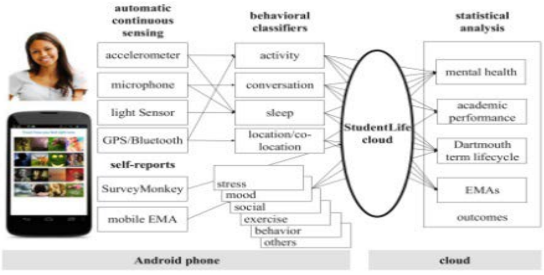


FIGURE 4. Dataset Structure and Collection Process [75].

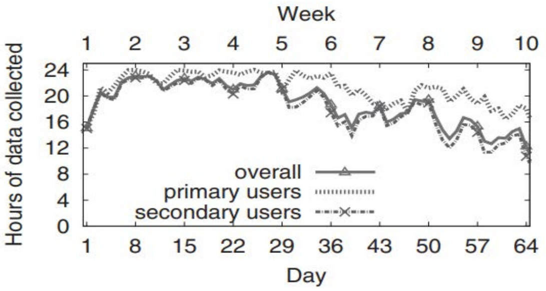


FIGURE 5. Numbers of Hours collecting Automatic sensing data over the term [75].

comma-separated values (CSV) files to assist in making the processing easier. Afterward, normalization, categorical data encoding, and removal of non-available data were applied. In addition to the standard pre-processing phase, time-series

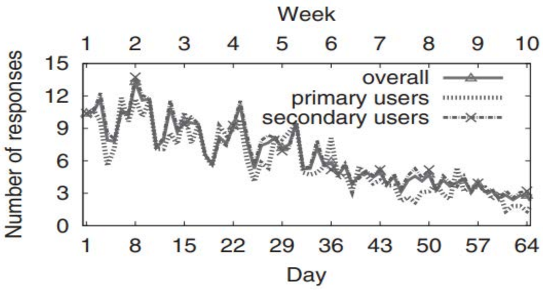


FIGURE 6. Number of users providing EMA data over the term [75].

and static clustering are experimented with before the regression phase.

Time series analysis (i.e., temporal data) increased significantly after realizing the importance of the time factor in generating better analytics. One of the main fields that have been influenced positively by the usage of time series data has been the education field. Most recently, the focus of detecting potential clusters between various people over time was applied in social network analysis [80].

In a previous study on the studentLife dataset [81], they were capable of analyzing various potential clusters to identify depression of students with an accuracy of 87%. Clusters were built on averaging features, checking changes (breakpoints), and location variance. These clusters considered only statistical analysis similarity, which does not consider the temporal changes within the students for time-series data. However, the technique in this previous research is still valid for static data, which will be repeated in the experiments for the current paper and compared with time-series clustering.

This work follows a 'cluster-then-predict' approach that was used previously to enhance prediction performance in various studies [82]–[84]. Cluster-then-predict is a methodology in which the observations are first clustered. Then cluster-specific prediction models are built per each cluster according to the students present in each cluster. However, clusters are formed here based on time-series data, not static data.

Initially, an exploratory visualization of Spatio-temporal features was conducted using Kepler.gl (a map based platform) to better understand how to divide the students into clusters according to their social behavior. This inspection assisted in understanding: 1- what the data represents, 2- what a cluster represents, 3- what the clustering is intended to achieve.

For the automated formation of clusters, the usage of multiple clustering techniques is examined with a combination of similarity measures, including k-means and hierarchical clustering, to recognize the patterns in the students' time-series based data (spatial or non-spatial). The performance of these algorithms is evaluated when time-series are multi-variate and of variable length. Multiple pre-processing functions are also applied to the time-series data before clustering operations to summarize each clustering algorithm's effectiveness with various forms of presented data.

For facilitating time-series handling, the paper utilized a library for time-series analysis called "tslearn" [85]. It is a python based library based on scikit-learn, which includes the most well-known algorithms and data analysis techniques for time-series data.

Time-series clustering can be achieved through the following two main steps:

- Measure similarity between students' Spatio-temporal data using Dynamic Time Wrapping (DTW) [86].
- Apply one of the standard clustering techniques (i.e., K-means [87]).

DTW works for computing similarity between temporal features of various students, Given series

$$\mathbf{X} = (\mathrm{x0}, \ldots, \mathrm{x_n})$$

and series

$$\mathbf{Y} = (\mathrm{yo}, \ldots, \mathrm{y_m}),$$

the DTW distance from X to Y is formulated as the following optimization problem:

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j)c\pi} d\left(x_i, y_j\right)^2} \quad (1)$$

where

$$\bar{\pi} = [\pi_0, \ldots, \pi_K]$$

is a path that satisfies the following properties:

- it is a list of index pairs $\pi_k = (i_k, j_k)$ with $0 \leq i_k < n$ and $0 \leq j_k < m$

- $\pi_0 = (0, 0)$ and $\pi_K = (n - 1, m - 1)$

- for all $k > 0, \pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:

$i_{k-1} \leq i_k \leq i_{k-1} + 1\, j_{k-1} \leq j_k \leq j_{k-1} + 1$

To summarize the DTW equation: DTW is calculated as the squared root of the sum of squared distances between each element in X and its nearest point in Y. Note that DTW(X,Y) not equal DTW(Y,X).

As a result, regardless of where temporal changes occur among the members, the centroids have an average form that resembles the shape of the cluster members.

Combining DTW with k-means steps is applied to cluster the students according to their similarities over time. For static features, only k-means is applied. In order to evaluate the clustering performed, specifically for multi-variate time series clustering, internal indexing is required. Internal indexing and internal validation are strategies that work without external data. These processes are used to process both classified and unclassified data collections. Indices include Dunn's index, Silhouette's, GK, Davies Bouldin (DBI), and Calinski Harabasz scores [88]. Following research on different forms of student patterns, the silhouette score is used [89], [90]. Silhouette score is a metric used to calculate the goodness of a clustering technique. The silhouette value measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Its value ranges from −1 to 1. Clusters that contain more grouped members and are separated cleanly from each other achieve a better silhouette score(0 and above). Details of these experiments are detailed in the experimental design section.

C. FEATURE SELECTION OF STUDENT DATA TO CoI CHARACTERISTICS

In [38], the principal investigators were assigned as subject matter experts. The current research utilized the same technique that was used for the transcript or text encoding. After individually coding the transcripts, the coders addressed the gaps between their coding via negotiation. This negotiation helped bring as much of their coding into an agreement as feasible. For each attribute or feature reviewed, agreement rates were calculated before and after negotiation. According to Garrison *et al.* [91], this negotiating process pushes inter-rater reliability into a condition of inter-subjectivity, when raters discuss, present, and dispute interpretations to see if agreement can be obtained.

In the paper in-hand, a similar methodology was applied for the Feature Selection. Three investigators were included in the coding process(between attribute names/descriptions and CoI indicators/descriptions). A negotiation agreement was utilized to reach a joint decision relying on the majority of the votes to settle the coding decision. The Selection was conducted with an agreement. This selection resulted in the reduction of 62.8% of the features, as many of the features were not representative of the student, under the umbrella of the CoI indicators. Sample of the coding conducted is presented in table 2.

D. GPA PREDICTION THROUGH REGRESSION

Regression-based models rely on solving an optimization problem in an attempt to minimize the mean square error from the expected model outcome [92]. One of the regression models used was LASSO regression, which automatically selects more relevant features and discards redundant features to avoid over-fitting [93].

A normal linear regression model is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \quad (2)$$

In order to enhance the performance of the regression, it is required to minimize the residual between the points. This residual is called the residual sum of squares (RSS). And by equation it is represented as:

$$RSS = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \right)^2 \quad (3)$$

where:

n: is the total number of observations (data).
y_i: is the actual output value of the observation (data).
p: is the total number of features.
x_{ij}: is a model's coefficient.
x_{ij}: is the ith observation, jth feature's value.

TABLE 2. Sample of the Selection between attributes and CoI framework.

Indicator/ Description	Tables (Including attributes/ Description)
Teaching Presence: Design & Organization. **Description:** The instructor clearly communicated important due dates/time frames for learning activities	EMA=>Stress **Description:** Ecological momentary assessment questionnaire. Includes rating of the attribute according to the student and exact response time and location.
Social Presence: Affective expression. **Description:** Online or web-based communication is an excellent medium for social interaction	Education =>piazza **Description:** Piazza usage statistics: Inlcuding number of topics, number of comments, and activities performed on the online platform.
Cognitive Presence: Triggering event. **Description:** I felt motivated to explore content related questions	EMA =>Study Spaces **Description:** Ecological momentary assessment questionnaire. Includes rating of the attribute according to the student and exact response time and location.

$\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$: is the predicted output of each observation.

In the case of using Ridge regression (or “L2 regularization”), It minimizes: RSS $+ \alpha \sum_{j=1}^{p} \beta_j^2$ However, for Lasso regression (or “L1 regularization”), it minimizes: RSS $+ \alpha \sum_{j=1}^{p} |\beta_j|$.

Other algorithms such as Huber regressor [94] utilize a combination of L1 and L2 regularization, even with a different loss function.

Various regression techniques have been employed by previous research for GPA prediction. This includes standard Linear, Lasso [93], LassoCV [95], and Ridge Regression techniques [96].

IV. EXPERIMENTAL DESIGN

This section describes the experiments conducted to ensure the methodology was performed correctly. The experiments are conducted after the initial data wrangling and feature selection according to the CoI Process. These experiments assisted in reaching the optimal parameters to conduct the final experiments for the GPA prediction. For the clustering phase, various temporal features were used, but no clusters were formed except the GPS feature. Details are mentioned in the subsection below.

A. EXPERIMENTS DETAILS

The first experiment conducted was the visualization of students' movement over the map in order to ensure if there is a specific pattern in their GPS locations or not. In figure 7, various samples of students' movement patterns over time are presented. As shown, there are some common patterns between the students. These patterns can be classified into three main categories:

- Students who overtime move locally within the same city (Figure 7(a)).
- Students who overtime move within the same state or within the East Coast (several states at the same side)(Figure 7(b)).
- Students who move all over the united states or globally(Figure 7(c)). This cluster can be separated into two different clusters if the number of students in each cluster increases.

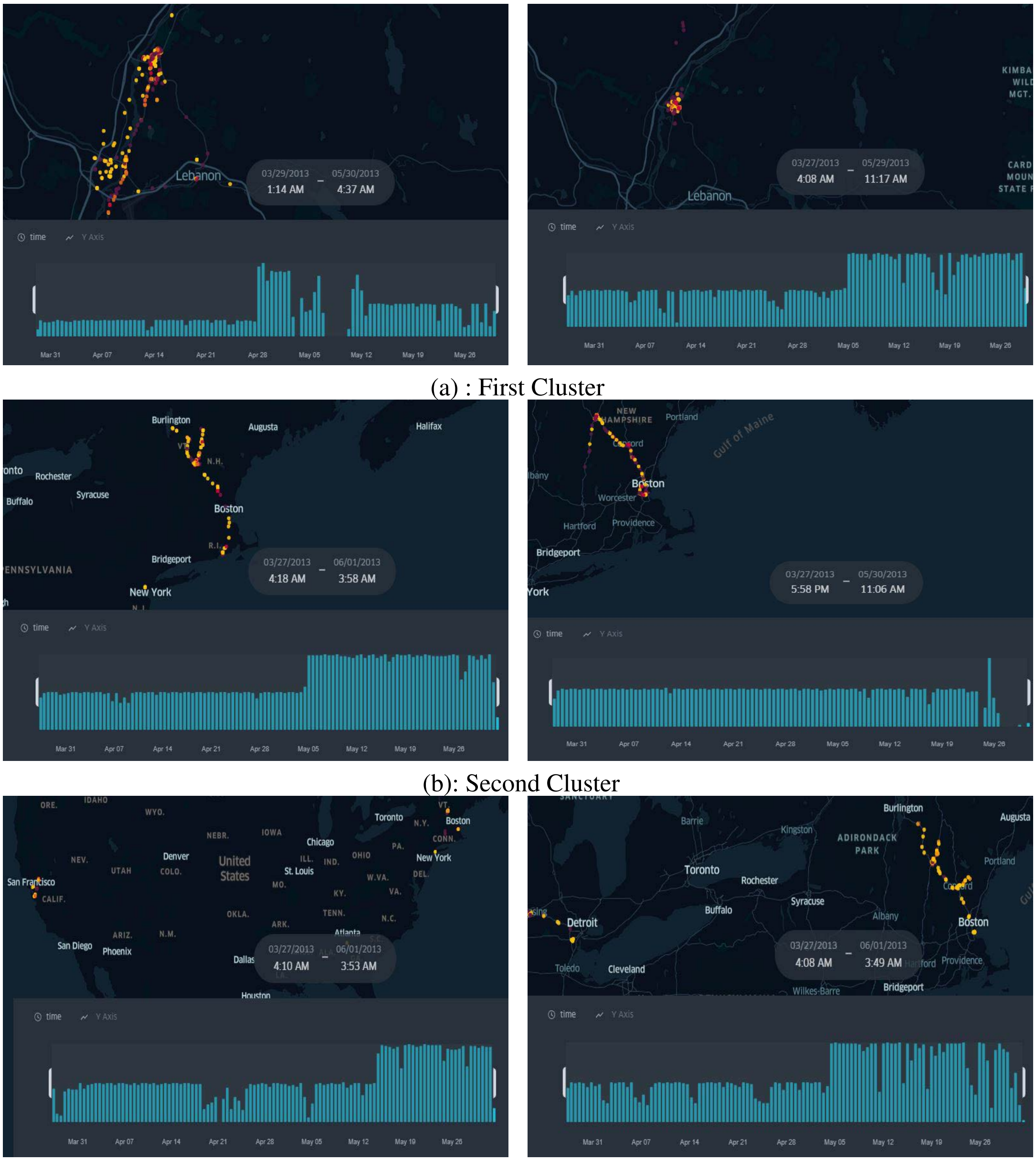
Following the visualization, as a second experiment, the automated process of clustering using DTW similarity and k-means was applied. The number of clusters chosen should achieve the best silhouette score, as shown in Figure 8. The number of clusters generated with the automated process agreed with the initial manual inspection. For comparison purposes, clustering over one of the static data available (The static feature is: PHQ-9 depression scale scores [97]). This feature was collected within pre-semester and post-semester surveys and averaged to form the clusters. The clusters were similar to the one generated in [81]. Cluster 1 of a normalized mean value of (-0.37), Cluster 2 of mean values (0.39), and Cluster 3 of a mean value of (0.05). As mentioned with time-series clustering, these clusters will be separated to generate various prediction models per each cluster.

After clustering, various regression models are applied for prediction. The results are detailed in the Results section.

B. EVALUATION

The evaluation of the results of the prediction process is performed according to multiple evaluation techniques. These techniques are detailed as follows:

- **MAE (Mean Absolute Error):** MAE assesses the average magnitude of mistakes in a set of forecasts without considering their direction. It is the average of the absolute differences between forecast and actual observation over the test sample, where all individual deviations are given equal weight. The lower the MAE, the better a model fits a dataset. MAE was used in previous research on the same dataset.
- **RMSE (Root mean square error)**: A metric that tells us how far apart the predicted values are from the observed values in a dataset, on average. The lower the RMSE, the better a model fits a dataset. The RMSE is a good



(a) : First Cluster

(b): Second Cluster

(c) : Third Cluster

FIGURE 7. Samples of Students Goe-locations overtime.

measure for evaluating the performance of a model because RMSE is proportional to the observed mean.

- **Adjusted R-squared:** R-Squared (R2) is a metric that tells us the proportion of the variance in the response variable of a regression model that the predictor variables can explain. This value ranges from 0 to 1. The higher the R2 value, the better a model fits a dataset. R2 is the square of the correlation coefficient, which is computed as follows: Correlation Coefficient (r):

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)(\Sigma y)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - (\Sigma y)^2\right]}} \quad (4)$$

where;

x is the observed variable, y is the predicted variable, and n are the number of observations.

One pitfall of R-squared is that the value of R2 can only increase as predictors are added to the regression model. This increase is artificial when predictors are not improving the model's fit. A related statistic, Adjusted R-squared, will decrease as predictors are added if the increase in model fit does not compensate for the loss of degrees of freedom.

Cross-Validation is applied to ensure the stability of the results. Cross-validation is a resampling method that uses different data portions to test and train a model on different iterations. Repeated k-fold cross-validation is utilized with

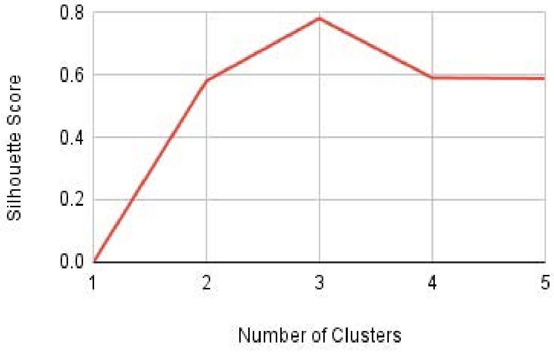


FIGURE 8. Silhouette Score Versus Number of Clusters.

values of: k (partitions=10), r(repeats)=3. This k-fold division has been utilized in earlier work [98]. Other techniques such as Monte Carlo simulations are utilized in previous research related to forecasting [99]. However, K-fold was selected as K-fold cross-validation is the most widely used method in social sciences [100].

In addition to the previous evaluation criteria, the runtime and number of features utilized in each run are detailed to compare efficiency. The platform on which these experiments were conducted was Google Colab, using their standard CPU-based runtime.

V. RESULTS AND DISCUSSION

The purpose of the research detailed in the paper is to examine combining social-based clustering techniques with various regressions. In addition, another set of experiments considers the feature selection based on the CoI model and how it influences the prediction performance. Table 3 summarizes the results achieved from the experiments.

A. GPA PREDICTION THROUGH REGRESSION (BASELINE)

1) ACCURACY

The first experiment aims to utilize regression techniques for GPA prediction.The experiment was conducted by applying the previous work [45], which applies different regression techniques for GPA prediction on the same dataset and compares it with other regressive techniques. The best results from previous research achieved 0.179 mean error, correlation (r)= 0.81 using LassoCV. The techniques used are standard Linear regression, Lasso Regression, LassoCV, and Ridge regression with variations in the algorithms' factors such as the Regularization parameters. As presented in the methodology section, multiple techniques for GPA prediction have been evaluated. Firstly, the comparison between the regression models has resulted in support for the usage of LassoCV for a better prediction output. In general, it can be perceived that algorithms that focus on the L1 regularization only (ex: Lasso, LassoCV) as a penalty perform better than other algorithms that include L2 regularization (ex: ridge). The reason would be that algorithms based on L1 regularization would eliminate the least essential features and the selection performed.

Note that the results per each trial/algorithm represent the average range of accuracies of each technique while using multiple regularization factors (alpha). The range of regularization values used is from 0.001 to 0.01. Higher values were tested but achieved near-zero accuracy as it eliminated most of the features necessary for the prediction process. This technique works effectively in datasets with a larger number of features. Standard Linear regression performed as expected of not converging with more features presented. As per figures 10, 11, 12, LassoCV outperforms other techniques in MA, RMSE, and Adjusted R2, specifically in the case of no clustering and no feature selection. More details about other experiments are mentioned in the below subsections.

2) RUNTIME

In the first trial (without feature selection according to the CoI model). The utilization of all set of features increased the potential runtime for all algorithms. Detailed comparison of all techniques is presented in figure 9. The runtime results agree with previous research [101], [102]. LassoCV has a higher runtime due to the cross-validation applied to select the best (alpha) for the regression model.

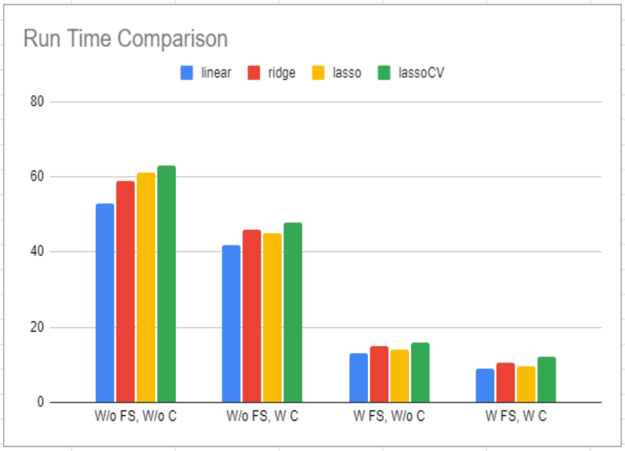


FIGURE 9. Runtime of algorithms (in Seconds).

B. FEATURE SELECTION OF STUDENT DATA TO CoI CHARACTERISTICS

1) ACCURACY

It was concluded from the previous section that LassoCV is considered the best performing regressive technique, with a trade-off in comparison to LASSO in terms of time due to the application of least angle cross-validation [103]. For the second phase of the work conducted, a trial is applied by using features selected according to the CoI model. This selection resulted in focusing only on the features most representative of the student according to the CoI framework. Accuracy enhancement was minor but still apparent. The number of features was reduced from 218 to 81, as shown in figure 13,

TABLE 3. Comparative Analysis Results (Accuracies).

Results				
	Evaluation Metric			
Technique	MAE	RMSE	Adjusted R2	Number of Features
Without Feature Selection (W/o FS)				
Without Clustering (W/o C)				
Linear	0.172	0.287	0.577	218
Ridge	0.189	0.203	0.4	218
Lasso	0.103	0.189	0.88	218
LassoCV	0.09	0.145	0.92	218
With Clustering (W C)				
Linear	0.193	0.26	0.59	218
Ridge	0.186	0.217	0.456	218
Lasso	0.17	0.201	0.842	218
LassoCV	0.11	0.187	0.91	218
With Feature Selection (W FS)				
Without Clustering (W/o C)				
Linear	0.204	0.296	0.6	81
Ridge	0.217	0.354	0.52	81
Lasso	0.156	0.162	0.883	81
LassoCV	0.126	0.15	0.95	81
With Clustering (W C)				
Linear	0.225	0.314	0.64	**81**
Ridge	0.293	0.423	0.296	**81**
Lasso	0.112	0.184	0.91	**81**
LassoCV	**0.07**	**0.135**	**0.96**	**81**

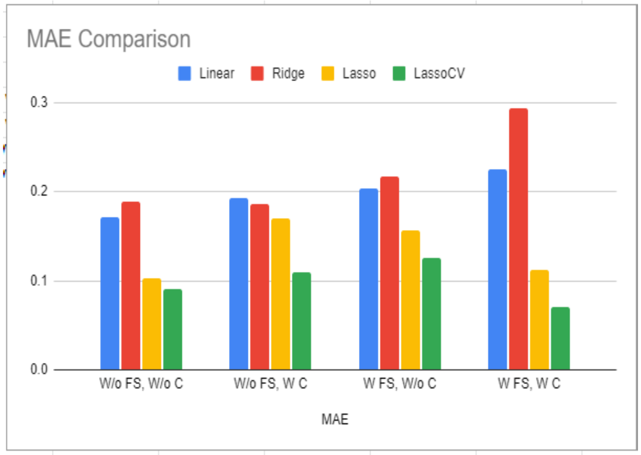


FIGURE 10. Mean Average Error Results.

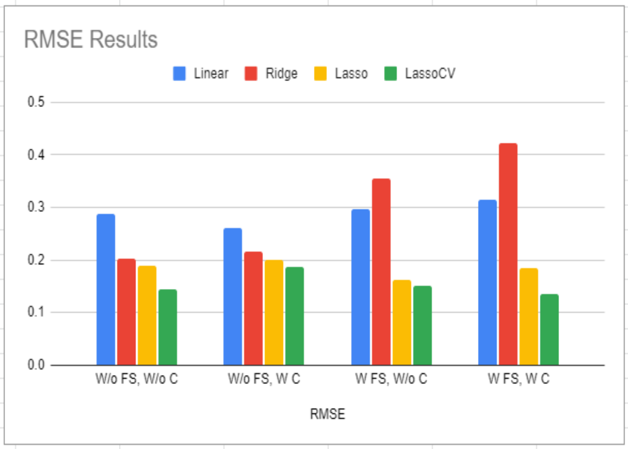


FIGURE 11. Root Mean Square Error Results.

and accordingly, the run-time was affected. It is to be noted that the performance of the features selection without clustering overcomes the performance of using clustering but without using the feature selection, meaning that selecting the attribute presenting the student ensures a better-situated model.

2) RUNTIME

Regarding the algorithms' runtime, it can be viewed from figure 9 that the runtime of the regression models is reduced. This reduction is due to utilizing a lower number of features as an input to the regression models, as occurred in previous research [104].

C. PRE-PREDICTION PROCESSING USING STATIC AND TIME-SERIES CLUSTERING

1) ACCURACY

The third phase involved adding a clustering phase before the prediction process. The 'cluster-then-predict' part achieved no significant improvement in accuracy (Figures 10, 11, 12). However, Clustering was seen to perform better with feature selection rather than on its own.

The clustering over static values (PHQ-9 feature) showed no significant improvement over the baseline results (in cases showed deterioration of RMSE above 0.4). However, this might be because the clusters formed in this case are based on averaging two values, which might not be a good representative of the state of depression at a specific time.

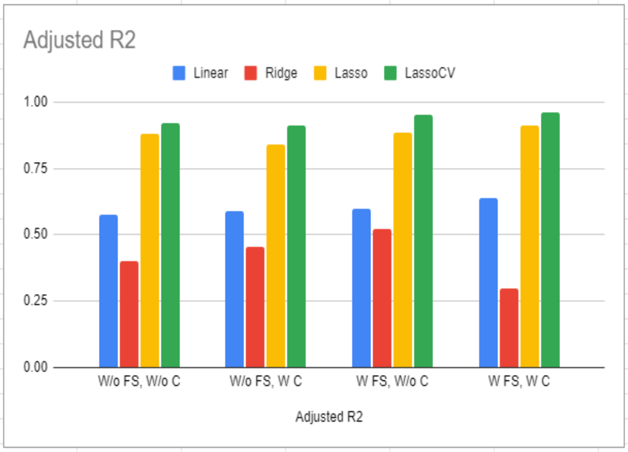


FIGURE 12. Adjusted R2 Results.

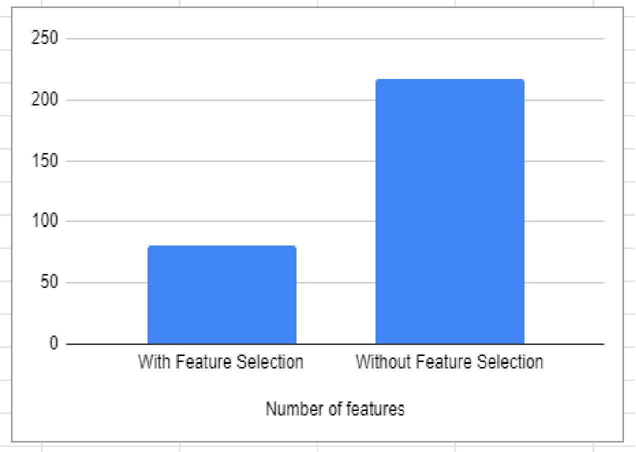


FIGURE 13. Number of Features.

2) RUNTIME

It was shown through the last phase that considering the temporal effects of social data (represented in the overtime movement of the student measured by the GPS location) can potentially reduce the runtime for the model if experimented on new students (Figure 9. The reduction in runtime matches previous research conclusions, in that smaller representative datasets achieved through clusters reduce regression processing time over larger datasets [105]. Futuristically, the runtime for new students will be reduced by identifying each new student's cluster. According to this cluster, they will only be classified, which reduces the amount of time used to rebuild the regression model. This result confirms the theory that a 'cluster-then-predict' technique can be helpful in the prediction of outcomes within the education field.

In conclusion, the results show performance precedence using LASSO and LASSOCV (with a trade-off regarding the runtime difference). This variation in runtime between various regression models. Time-Series clustering on its did not achieve the best performance; however, combined with the feature selection based on CoI, it achieved better performance in accuracy and runtime.

VI. CONCLUSION AND FUTURE WORK

In this paper, a comprehensive study has been implemented in the field of educational performance analysis. This work was divided into two main objectives:

- Analyzing the suitability of Community of Inquiry framework as a student modelling technique for selecting the attributes most representative of the students. This analysis is conducted by selecting students' attributes that fit the CoI model, a model that considers both cognitive and social aspects of the student. A comparison is then performed between the GPA prediction results and the results achieved by predicting through all available attributes to validate the suitability of the feature selection.
- Evaluating how the addition of a pre-prediction phase using time-series clustering over time-series based attributes can minimize the number of attributes required in the prediction process.

The results have shown that the techniques that were preceded by feature selection according to the CoI model, or adding a pre-prediction phase using clustering can enhance the prediction accuracy and minimize the required number of attributes. These results also open the door for how structured behavioral and educational models can be combined with automatically measured data to understand better how each attribute contributes to the evaluation of student performance. In addition, this research facilitates opportunities for future research in several areas. These areas include but not limited to:

- Automating the selection process between students' attributes and a standardized framework such as the CoI model utilized in this research. This automation can be performed through various sentence-relatedness techniques between attribute names and CoI indicators [106].
- Automating the process of identifying the number of clusters through hyper-parameters optimization. In addition, other non-linear techniques such as deep learning might be usable in the case of a lower number of features dataset [107].
- Studying how the CoI model may assist in other areas through the educational process other than student outcome prediction. This study may include how the CoI framework can automatically assist in generating courses suitable to various students based on their CoI presences.

REFERENCES

[1] R. Summers, H. Higson, and E. Moores, "The impact of disadvantage on higher education engagement during different delivery modes: A pre- versus peri-pandemic comparison of learning analytics data," *Assessment Eval. Higher Educ.*, pp. 1–11, Jan. 2022.

[2] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer, "Mining big data in education: Affordances and challenges," *Rev. Res. Educ.*, vol. 44, no. 1, pp. 130–160, Mar. 2020.

[3] T. Byers, W. Imms, and E. Hartnell-Young, "Making the case for space: The effect of learning spaces on teaching and learning," *Curriculum Teaching*, vol. 29, no. 1, pp. 5–19, Jan. 2014.

[4] P. A. Kirschner and J. J. G. van Merriënboer, "Do learners really know best? Urban legends in education," *Educ. Psychol.*, vol. 48, no. 3, pp. 169–183, Jul. 2013.

[5] L. Dooley and N. Makasis, "Understanding student behavior in a flipped classroom: Interpreting learning analytics data in the veterinary pre-clinical sciences," *Educ. Sci.*, vol. 10, no. 10, p. 260, Sep. 2020.

[6] D. Kurniadi, E. Abdurachman, H. Warnars, and W. Suparta, "A proposed framework in an intelligent recommender system for the college student," *J. Phys., Conf. Ser.*, vol. 1402, no. 6, 2019, Art. no. 066100.

[7] M. Marienko, Y. Nosenko, A. Sukhikh, V. Tataurov, and M. Shyshkina, "Personalization of learning through adaptive technologies in the context of sustainable development of teachers education," 2020, *arXiv:2006.05810*.

[8] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827–203844, 2020.

[9] M. A. A. Mamun, G. Lawrie, and T. Wright, "Instructional design of scaffolded online learning modules for self-directed and inquiry-based learning environments," *Comput. Educ.*, vol. 144, Jan. 2020, Art. no. 103695.

[10] K. A. Walker and K. E. Koralesky, "Student and instructor perceptions of engagement after the rapid online transition of teaching due to COVID-19," *Natural Sci. Educ.*, vol. 50, no. 1, Jan. 2021.

[11] I. Boticki, G. Akçapınar, and H. Ogata, "E-book user modelling through learning analytics: The case of learner engagement and reading styles," *Interact. Learn. Environments*, vol. 27, nos. 5–6, pp. 754–765, Aug. 2019.

[12] S. Sosnovsky, L. Müter, M. Valkenier, M. Brinkhuis, and A. Hofman, "Detection of student modelling anomalies," in *Proc. Eur. Conf. Technol. Enhanced Learn.* Cham, Switzerland: Springer, 2018, pp. 531–536.

[13] J. Kuzilek, J. Vaclavek, V. Fuglik, and Z. Zdrahal, "Student drop-out modelling using virtual learning environment behaviour data," in *Proc. Eur. Conf. Technol. Enhanced Learn.* Cham, Switzerland: Springer, 2018, pp. 166–171.

[14] B. Farr-Wharton, M. B. Charles, R. Keast, G. Woolcott, and D. Chamberlain, "Why lecturers still matter: The impact of lecturer-student exchange on student engagement and intention to leave university prematurely," *Higher Educ.*, vol. 75, no. 1, pp. 167–185, Jan. 2018.

[15] M. A. Islam, S. D. Barna, H. Raihan, M. N. A. Khan, and M. T. Hossain, "Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based cross-sectional survey," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0238162.

[16] J. A. Kumar and B. Bervell, "Google classroom for mobile learning in higher education: Modelling the initial perceptions of students," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1793–1817, Mar. 2019.

[17] S. Vhaduri, S. V. Dibbo, and Y. Kim, "Deriving college students' phone call patterns to improve student life," *IEEE Access*, vol. 9, pp. 96453–96465, 2021.

[18] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Proc. Comput. Sci.*, vol. 72, pp. 414–422, Dec. 2015.

[19] R. A. W. Tortorella and S. Graf, "Considering learning styles and context-awareness for mobile adaptive learning," *Educ. Inf. Technol.*, vol. 22, no. 1, pp. 297–315, 2017.

[20] M. Faisal, A. Bourahma, and F. AlShahwan, "Towards a reference model for sensor-supported learning systems," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 9, pp. 1145–1157, Nov. 2021.

[21] R. A. Sottilare, A. Graesser, X. Hu, and H. Holden, *Design Recommendations for Intelligent Tutoring Systems: Volume 1-Learner Modeling*, vol. 1. U.S. Army Research Laboratory, 2013.

[22] A. Abyaa, M. Khalidi Idrissi, and S. Bennani, "Learner modelling: Systematic review of the literature from the last 5 years," *Educ. Technol. Res. Develop.*, vol. 67, no. 5, pp. 1105–1143, Oct. 2019.

[23] K. Sharma, Z. Papamitsiou, and M. N. Giannakos, "Modelling learners' behaviour: A novel approach using garch with multimodal data," in *Proc. Eur. Conf. Technol. Enhanced Learn.* Cham, Switzerland: Springer, 2019, pp. 450–465.

[24] J. Verhagen, D. Hatfield, and D. Arena, "Toward a scalable learning analytics solution," in *Proc. Int. Conf. Artif. Intell. Educ.* Cham, Switzerland: Springer, 2019, pp. 404–408.

[25] P. Jiang and X. Wang, "Preference cognitive diagnosis for student performance prediction," *IEEE Access*, vol. 8, pp. 219775–219787, 2020.

[26] S. Järvelä, J. Malmberg, E. Haataja, M. Sobocinski, and P. A. Kirschner, "What multimodal data can tell us about the students' regulation of their learning process?" *Learn. Instruct.*, vol. 72, Apr. 2021, Art. no. 101203.

[27] J. Bicans and J. Grundspenkis, "Student learning style extraction from on-campus learning context data," *Proc. Comput. Sci.*, vol. 104, pp. 272–278, Jan. 2017.

[28] G.-C. Kim and R. Gurvitch, "Online education research adopting the community of inquiry framework: A systematic review," *Quest*, vol. 72, no. 4, pp. 395–409, Oct. 2020.

[29] D. R. Garrison, T. Anderson, and W. Archer, "A theory of critical inquiry in online distance education," *Handbook Distance Educ.*, vol. 1, no. 4, pp. 113–127, 2003.

[30] D. Yeats, "In a nutshell: Community of inquiry for online subjects," Tech. Rep., Nov. 2020.

[31] J. Valverde-Berrocoso, M. D. Garrido-Arroyo, C. Burgos-Videla, and M. B. Morales-Cevallos, "Trends in educational research about e-learning: A systematic literature review (2009–2018)," *Sustainability*, vol. 12, no. 12, p. 5153, 2020.

[32] B. R. Griffiths, "Student engagement in an online graduate business program and academic achievement," Ph.D. dissertation, Temple Univ., Philadelphia, PA, USA, 2020.

[33] Y. H. Lee and K.-J. Kim, "Enhancement of student perceptions of learner-centeredness and community of inquiry in flipped classrooms," *BMC Med. Educ.*, vol. 18, no. 1, pp. 1–6, Dec. 2018.

[34] T. W. McClannon, A. Cheney, L. Bolt, and K. Terry, "Predicting sense of presence and sense of community in immersive online learning environments," *Online Learn.*, vol. 22, no. 4, pp. 141–159, Dec. 2018.

[35] M. Ferreira, V. Rolim, R. F. Mello, R. D. Lins, G. Chen, and D. Gašević, "Towards automatic content analysis of social presence in transcripts of online discussions," in *Proc. 10th Int. Conf. Learn. Anal. Knowl.*, Mar. 2020, pp. 141–150.

[36] H. Hayati, A. Chanaa, M. K. Idrissi, and S. Bennani, "Doc2Vec &Naïve Bayes: Learners' cognitive presence assessment through asynchronous online discussion TQ transcripts," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 8, pp. 1–12, 2019.

[37] E. Ammenwerth, M. Netzer, and W. O. Hackl, "Learning analytics and the community of inquiry: Indicators to analyze and visualize online-based learning," in *dHealth*, 2020, pp. 67–68.

[38] M. Kaul, M. Aksela, and X. Wu, "Dynamics of the community of inquiry (CoI) within a massive open online course (MOOC) for in-service teachers in environmental education," *Educ. Sci.*, vol. 8, no. 2, p. 40, Mar. 2018.

[39] K. Kozan and S. Caskurlu, "On the Nth presence for the community of inquiry framework," *Comput. Educ.*, vol. 122, pp. 104–118, Jul. 2018.

[40] W. Boston, S. R. Díaz, A. M. Gibson, P. Ice, J. Richardson, and K. Swan, "An exploration of the relationship between indicators of the community of inquiry framework and retention in online programs," Tech. Rep., 2009.

[41] K. Sheridan and M. A. Kelly, "The indicators of instructor presence that are important to students in online courses," *J. Online Learn. Teach.*, vol. 6, no. 4, p. 767, 2010.

[42] S. Joksimovic, D. Gasevic, V. Kovanovic, O. Adesope, and M. Hatala, "Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions," *Internet Higher Educ.*, vol. 22, pp. 1–10, Jul. 2014.

[43] P. Guo, N. Saab, L. Wu, and W. Admiraal, "The community of inquiry perspective on students' social presence, cognitive presence, and academic performance in online project-based learning," *J. Comput. Assist. Learn.*, vol. 37, no. 5, pp. 1479–1493, Oct. 2021.

[44] A. A. Saa, "Educational data mining & students' performance prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 212–220, 2016.

[45] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2014, pp. 3–14.

[46] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell, "SmartGPA: How smartphones can assess and predict academic performance of college students," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2015, pp. 295–306.

[47] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 1913–1927, May 2020.

[48] J. Xu, K. H. Moon, and M. van der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 5, pp. 742–753, Aug. 2017.

[49] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 415–421.

[50] B. Rubin, R. Fernandes, and M. D. Avgerinou, "The effects of technology on the community of inquiry and satisfaction with online courses," *Internet Higher Educ.*, vol. 17, pp. 48–57, Apr. 2013.

[51] V. Kovanović, D. Gašević, S. Joksimović, M. Hatala, and O. Adesope, "Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions," *Internet Higher Educ.*, vol. 27, pp. 74–89, Oct. 2015.

[52] V. Kovanovic, "Assessing cognitive presence using automated learning analytics methods," Tech. Rep., 2017.

[53] V. Kovanović, S. Joksimović, O. Poquet, T. Hennis, P. de Vries, M. Hatala, S. Dawson, G. Siemens, and D. Gašević, "Examining communities of inquiry in massive open online courses: The role of study strategies," *Internet Higher Educ.*, vol. 40, pp. 20–43, Jan. 2019.

[54] J. L. F. Choy and C. L. Quek, "Modelling relationships between students' academic achievement and community of inquiry in an online learning environment for a blended course," *Australas. J. Educ. Technol.*, vol. 32, no. 4, 2016.

[55] J. T. Abbitt and W. J. Boone, "Gaining insight from survey data: An analysis of the community of inquiry survey using Rasch measurement techniques," *J. Comput. Higher Educ.*, vol. 33, no. 2, pp. 367–397, 2021.

[56] S. Xu, H. Luo, and Y. Tan, "Re-examining the community of inquiry framework from the perspective of learning analytics," in *Proc. 13th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2018, pp. 1–5.

[57] W. Purba, S. Tamba, and J. Saragih, "The effect of mining data k-means clustering toward students profile model drop out potential," *J. Phys., Conf. Ser.*, vol. 1007, Apr. 2018, Art. no. 012049.

[58] R. G. Santosa, Y. Lukito, and A. R. Chrismanto, "Classification and prediction of students' GPA using K-means clustering algorithm to assist student admission process," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 7, no. 1, pp. 1–10, 2021.

[59] H. Turabieh, S. A. Azwari, M. Rokaya, W. Alosaimi, A. Alharbi, W. Alhakami, and M. Alnfiai, "Enhanced Harris hawks optimization as a feature selection for the prediction of student performance," *Computing*, pp. 1417–1438, Jan. 2021.

[60] R. Soni and K. J. Mathai, "Improved Twitter sentiment prediction through cluster-then-predict model," 2015, *arXiv:1509.02437*.

[61] E. Popescu and G. Badea, "Exploring a community of inquiry supported by a social media-based learning environment," *Educ. Technol. Soc.*, vol. 23, no. 2, pp. 61–76, 2020.

[62] P. Chappell, M. Tse, M. Zhang, and S. Moore, "Using GPS geo-tagged social media data and geodemographics to investigate social differences: A Twitter pilot study," *Sociol. Res. Online*, vol. 22, no. 3, pp. 38–56, Sep. 2017.

[63] Z. Li, X. Huang, X. Ye, Y. Jiang, M. Yago, H. Ning, M. E. Hodgson, and X. Li, "Measuring global multi-scale place connectivity using geotagged social media data," 2021, *arXiv:2102.03991*.

[64] D. R. Garrison, "Online community of inquiry review: Social, cognitive, and teaching presence issues," *J. Asynchronous Learn. Netw.*, vol. 11, no. 1, pp. 61–72, 2007.

[65] N. Pellas and A. Boumpa, "Open sim and sloodle integration for preservice foreign language teachers' continuing professional development: A comparative analysis of learning effectiveness using the community of inquiry model," *J. Educ. Comput. Res.*, vol. 54, no. 3, pp. 407–440, Jun. 2016.

[66] P. Ice, A. M. Gibson, W. Boston, and D. Becher, "An exploration of differences between community of inquiry indicators in low and high disenrollment online courses," *J. Asynchronous Learn. Netw.*, vol. 15, no. 2, pp. 44–69, 2011.

[67] P. Shea, S. Hayes, and J. Vickers, "Online instructional effort measured through the lens of teaching presence in the community of inquiry framework: A re-examination of measures and approach," *Int. Rev. Res. Open Distrib. Learn.*, vol. 11, no. 3, pp. 127–154, 2010.

[68] E. Keles, "Use of Facebook for the community services practices course: Community of inquiry as a theoretical framework," *Comput. Educ.*, vol. 116, pp. 203–224, Jan. 2018.

[69] D. Dell, "Emotional presence in community of inquiry: A scoping review and delphi study," Tech. Rep., 2021.

[70] A. T. Tolu, *An Exploration of Synchronous Communication in an Online Preservice ESOL Course: Community of Inquiry Perspective*. Tampa, FL, USA: Univ. South Florida, 2010.

[71] *Large-Scale WebGL-Powered Geospatial Data Visualization Tool*, Uber, San Francisco, CA, USA, 2020.

[72] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Sci. Data*, vol. 4, no. 1, pp. 1–8, Dec. 2017.

[73] S. Dietze, D. Taibi, and M. d'Aquin, "Facilitating scientometrics in learning analytics and educational data mining—The LAK dataset," *Semantic Web*, vol. 8, no. 3, pp. 395–403, Dec. 2016.

[74] U. Dartmouth, "Studentlife dataset," Tech. Rep., 2020.

[75] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Using smartphones to assess mental health and academic performance of college students," in *Mobile Health*. Cham, Switzerland: Springer, 2017, pp. 7–33.

[76] G. Mikelsons, M. Smith, A. Mehrotra, and M. Musolesi, "Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities," 2017, *arXiv:1711.06350*.

[77] T. Tominaga, S. Yamamoto, T. Kurashima, and H. Toda, "Effects of personal characteristics on temporal response patterns in ecological momentary assessments," in *Proc. IFIP Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2021, pp. 3–22.

[78] B. Nguyen, S. Kolappan, V. Bhat, and S. Krishnan, "Clustering and feature analysis of smartphone data for depression monitoring," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 113–116.

[79] Q. He and E. O. Agu, "Context-aware probabilistic models for predicting future sedentary behaviors of smartphone users," *J. Healthcare Informat. Res.*, vol. 6, no. 1, pp. 112–152, Mar. 2022.

[80] N. Alotaibi and D. Rhouma, "A review on community structures detection in time evolving social networks," *J. King Saud Univ.-Comput. Inf. Sci.*, Aug. 2021.

[81] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis, "Multi-view bi-clustering to identify smartphone sensing features indicative of depression," in *Proc. IEEE 1st Int. Conf. Connected Health, Appl., Syst. Eng. Technol. (CHASE)*, Jun. 2016, pp. 264–273.

[82] N. K. Nagwani and S. V. Deo, "Estimating the concrete compressive strength using hard clustering and fuzzy clustering based regression techniques," *Sci. World J.*, vol. 2014, pp. 1–16, Oct. 2014.

[83] R. Soni and K. J. Mathai, "An innovative 'cluster-then-predict' approach for improved sentiment prediction," in *Advanced Computing and Communication Technologies*. Singapore: Springer, 2016, pp. 131–140.

[84] X. Song, W. Li, D. Ma, D. Wang, L. Qu, and Y. Wang, "A match-then-predict method for daily traffic flow forecasting based on group method of data handling," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 11, pp. 982–998, Nov. 2018.

[85] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, a machine learning toolkit for time series data," *J. Mach. Learn. Res.*, vol. 21, no. 118, pp. 1–6, 2020.

[86] T. Giorgino, "Computing and visualizing dynamic time warping alignments in R: The DTW package," *J. Stat. Softw.*, vol. 31, no. 7, pp. 1–24, 2009.

[87] X. Huang, Y. Ye, L. Xiong, R. Y. K. Lau, N. Jiang, and S. Wang, "Time series k-means: A new k-means type smooth subspace clustering for time series data," *Inf. Sci.*, vols. 367–368, pp. 1–13, Nov. 2016.

[88] T. Das, S. Paitnaik, and S. P. Mishra, "Identification of the optimal number of clusters in textual data," in *Advances in Distributed Computing and Machine Learning*. Singapore: Springer, 2022, pp. 215–225.

[89] J. Yin and M. M. Gaber, "Clustering distributed time series in sensor networks," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 678–687.

[90] Y.-A. Shih, B. Chang, and J. Y. Chin, "Data-driven student homophily pattern analysis of online discussion in a social network learning environment," *J. Comput. Educ.*, vol. 7, no. 3, pp. 373–394, Sep. 2020.

[91] D. R. Garrison, M. Cleveland-Innes, M. Koole, and J. Kappelman, "Revisiting methodological issues in transcript analysis: Negotiated coding and reliability," *Internet Higher Educ.*, vol. 9, no. 1, pp. 1–8, Jan. 2006.

[92] A. O. Sykes, "An introduction to regression analysis," Tech. Rep., 1993.

[93] J. Ranstam and J. Cook, "Lasso regression," Tech. Rep., 2018.

[94] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemp. Math.*, vol. 443, no. 7, pp. 59–72, 2007.

[95] H. Wang, G. Li, and C.-L. Tsai, "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 69, no. 1, pp. 63–78, Feb. 2007.

[96] G. C. McDonald, "Ridge regression," *Wiley Interdiscipl. Rev. Comput. Stat.*, vol. 1, no. 1, pp. 93–100, 2009.

[97] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Internal Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[98] K. T. Chui, R. W. Liu, M. Zhao, and P. O. De Pablos, "Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine," *IEEE Access*, vol. 8, pp. 86745–86752, 2020.

[99] N. D. Bokde, Z. M. Yaseen, and G. B. Andersen, "ForecastTB—An R package as a test-bench for time series forecasting—Application of wind speed and solar radiation modeling," *Energies*, vol. 13, no. 10, p. 2578, May 2020.

[100] M. D. Verhagen, "A pragmatist's guide to using prediction in the social sciences," *Socius*, vol. 8, pp. 1–17, Feb. 2022.

[101] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, p. 1, 2010.

[102] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. London, U.K.: Pearson, 2016.

[103] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[104] M. De Cock, R. Dowsley, A. C. A. Nascimento, D. Railsback, J. Shen, and A. Todoki, "High performance logistic regression for privacy-preserving genome analysis," *BMC Med. Genomics*, vol. 14, no. 1, pp. 1–18, Dec. 2021.

[105] A. Santana, S. Inoue, K. Murakami, T. Iizaka, and T. Matsui, "Clustering-based data reduction approach to speed up SVM in classification and regression tasks," in *Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.* Cham, Switzerland: Springer, 2020, pp. 478–488.

[106] M. Gonsalves, R. L. Whittles, R. B. Weisberg, and C. Beard, "A systematic review of the word sentence association paradigm (WSAP)," *J. Behav. Therapy Exp. Psychiatry*, vol. 64, pp. 133–148, Sep. 2019.

[107] A. Roy, M. Rahman, M. N. Islam, N. I. Saimon, M. Alfaz, and A.-A.-S. Jaber, "A deep learning approach to predict academic result and recommend study plan for improving student's academic performance," in *Ubiquitous Intelligent Systems*. Singapore: Springer, 2022, pp. 253–266.

YOMNA M. I. HASSAN received the master's degree in computing and information sciences from the Faculty of Engineering, Ain Shams University, in 2011, in a collaboration program organized through MIT. She is currently pursuing the Ph.D. degree with the Faculty of Artificial Intelligence and Computers, Cairo University, with a focus on student profiling and behavioral analysis.

She has over ten years of experience in machine learning and optimization applications, both within academic and industry-based entities. She has worked in multinational organizations where she was involved in multiple research projects producing patents and international publications. She currently works as an Assistant Lecturer (acting as a Lecturer) at Misr International University. She supervises graduation projects, prepares curriculum, and teaching courses, such as computer graphics, software engineering, and human–computer interaction. She is a reviewer for multiple international journals and conferences.

ABEER ELKORANY received the B.S., M.Sc., and Ph.D. degrees (Hons.) in electronics and communications engineering (EE) from the Faculty of Engineering, Cairo University, Egypt, in 1992, May 1996, and February 2002, respectively.

From 1993 to 2003, she was a Key Researcher in the expert system development at the Center Laboratory of Agriculture Expert System (CLAES), before joining the Department of Computer Science, Faculty of Computers and Information, Cairo University. She has taught many courses such as structure programming, object oriented programming, operating system concepts, advanced operating system, knowledge-based system, artificial intelligence, and semantic web. From March 2008 to December 2010, she worked as a Consultant with the Research and Innovation Support Section, Information Technology Industry Development Agency (ITIDA). From 2011 to 2015, she also worked as the Director of the Alumni Unit, and was the Founder of FCI Innovators Club, Faculty of Computers and Artificial Intelligence. She has authored/coauthored more than 50 publications in local and international periodicals and conferences, which qualified her to obtain the scientific publication award many times from the University of Cairo. Her research interests include quality assurance and knowledge-based system measurement, semantic networks, anthology development, knowledge management, social network analysis, and recommendation systems.

KHALED WASSIF received the B.Sc. degree (Hons.) in accounting from the Faculty of Commerce, Cairo University, in 1983, the P.G. Diploma degree in computer and information sciences from the Institute of Statistical Studies and Research, Cairo University, in 1986, and the master's and Ph.D. degrees in artificial intelligence from Cairo University, in 1991 and 1998, respectively.

He is currently an Associate Professor with the Faculty of Computers and Artificial Intelligence, Cairo University. He has supervised or co-supervised 12 students on their Ph.D. dissertations and M.S. theses. He has published 21 research papers in international journals and conference proceedings. His research interests include machine learning, data mining, web mining, case-based reasoning, big data, and soft computing. He is a reviewer in the international *Egyptian Informatics Journal* and the *Egyptian Computer Journal*.