Applying Machine Learning Techniques on Self-Reported Engagement and Student Log Data to Predict CS Learning Performance

1st Sultanah Abdullah A Albakri
University of Glasgow
Glasgow, UK
s.albakri.1@research.gla.ac.uk
University of Hail
Hail, Saudi Arabia
s.albakry@uoh.edu.sa

2nd Mireilla Bikanga Ada
University of Glasgow
Glasgow, UK
mireilla.bikangaada@glasgow.ac.uk

3rd Alistair Morrison
University of Glasgow
Glasgow, UK
alistair.morrison@glasgow.ac.uk

***Abstract*—Enhancing student engagement in computer science (CS) courses is crucial to improving students' achievement and fostering active participation in computer science education (CSE). Previous studies have highlighted different factors that shape student engagement, including behavioural, cognitive, emotional, and social engagement. Additionally, other factors influence student engagement, such as students' beliefs in the usefulness of learning computer science and their confidence in taking CS classes. Despite existing studies investigating student engagement in CSE, limited studies have explored factors that influence novice student engagement in CS courses. Further, no study has applied machine learning (ML) techniques on the combined self-reported engagement data and student log data to predict CS learning performance. Therefore, this study used ML techniques to explore and identify student engagement factors that affect and predict CS learning outcomes. To achieve this, data was collected using three different sources: self-reported data, system log data, and CS learning performance data. Log data from student behaviour on the system included monthly logs of student access to the CS course page on the LMS during the semester, the total hits of student interactions with the course content throughout the semester, and the total number of task submissions. Our analysis involves 77 novice students who consented and completed a multidimensional self-reported questionnaire during the second semester of 2022 – 2023 at a university in Saudi Arabia. The K-means clustering algorithm was used to understand engagement patterns by classifying students into groups based on their levels of self-reported engagement, log data, and academic performance. Classification algorithms using Random Forest (RF), Decision Tree (DT), and LightGBM (LGBM) were used to predict CS learning performance from student engagement data (self-reported and logs). We evaluated the performance of ML algorithms using metrics including accuracy, precision, recall, and F1-score. The clustering results showed that students who actively engage with the course content (log data) tend to achieve higher grades, especially those with higher total hits of student interactions with the course content throughout the semester. The classification results showed that the RF model outperforms DT and LGBM, highlighting the significance of student interactions with the course in the first month of the semester as the key indicator influencing CS learning performance. Our findings contribute to a deeper understanding of student engagement in CS education and highlight various sources and factors used to measure and influence student engagement. The study has implications for educators, researchers, and stakeholders who may design effective interventions that would increase engagement to improve student learning outcomes in CS education. Future work will use a larger sample of participants from various educational levels in different countries.**

***Index Terms*—Student Engagement, Log Data, Predict, CS Learning Performance**

I. Introduction

The number of students in the CS field increases each year due to its importance and relevance to all other disciplines in the modern era. However, novice students face many challenges in studying CS courses, which impact their retention in CS. One of these challenges is the difficulty of engagement in learning CS courses, which increases the difficulty of learning programming, as it is a complex process that requires different skills. Due to student engagement problems, the student's learning difficulties may increase; this negatively affects their achievements, interests, and effective reactions [1].

To help students overcome the challenges they face at the beginning of their learning journey in CS courses, it is important to help them increase their engagement with CS learning. We already know that positive student engagement has a positive relationship with CS learning outcomes [2]–[5]. However, we don't know which specific student engagement factors affect CS learning outcomes among novice students.

Previous research in CS education has explored student engagement, but it has made limited progress in identifying specific student engagement factors that predict CS learning outcomes, specifically CS learning performance. This is partly due to the complex nature of student engagement, which involves multiple constructs [6], [7]. In our prior research [8], [9], we have explored factors that shape and influence student engagement within CS classes. We investigated key factors, including behavioural, cognitive, emotional, and social dimensions, that shape and affect student engagement in CS education. We also explored two factors, including student

confidence and the perceived usefulness of CS, that have a significant positive influence on engagement levels, particularly for demographics considered underrepresented in the CS field, such as female students [8], [9]. We found that student engagement in CS depends on a wide variety of complex and interrelated factors (e.g., cognitive, and non-cognitive factors). However, the relationships between these influencing factors and CS learning performance have not yet been explored. This prompts a crucial question: Can we effectively predict student learning performance based on factors that shape and influence their engagement?

This research problem requires in-depth investigation to identify the key student engagement factors that affect and predict CS learning performance in CS education. Given the importance of student engagement in CS, it is crucial to investigate and measure student engagement factors using diverse methods among novice students in CS classes. These factors are complicated in terms of their depth, making it challenging to use standard quantitative analyses [10] where often, research is reduced to only one or a few essential factors [10], [11]. Therefore, this study builds on our previous work [8], [9], [12], aiming to use ML to provide a deeper understanding of the relationship between student engagement dimensions, confidence, perceived usefulness, and academic performance within CSE for novice students in the higher education (HE).

II. BACKGROUND

A. Student Engagement

This study builds on the self-system motivation theory [13], which suggests that engagement in learning is influenced by how students interact with their environment and how contextual characteristics vary [7]. The quality of the learning activity provides students with insights into their competence, their relationships with others in the learning environment, and their autonomy as learners. These insights accumulate over time, shaping students' engagement in various educational activities and influencing their learning outcomes [7]. Using the self-system motivation theory, engagement in this study refers to both observable and hidden factors of how students engage with learning activities in CS classes. We specifically examined four dimensions of engagement : behavioural, emotional, cognitive, and social engagement [9], which have been adapted from a multidimensional student engagement framework [6], [7]. This framework provides a more comprehensive and nuanced understanding of the factors that contribute to student engagement, including the individual and contextual factors that influence different types of engagement [6] and can inform the development of effective strategies to promote student engagement in various educational settings [7].

Behavioural engagement (BE) describes the student's behaviours in the class, such as attendance, participation, and completion of assignments. Cognitive engagement (CE) focuses on the mental processes involved in learning, such as critical thinking, and problem-solving. Emotional engagement (EE) refers to the feelings and emotions students experience in the class, such as interest, and enjoyment. Social engagement (SE) focuses on the interpersonal relationships that students develop with teachers, peers, and other members of the learning community [6]–[9], [12].

Further, the literature highlights two important factors for engagement in CSE: student confidence in CS learning [9], [14] and their beliefs in the usefulness of learning CS (perceived usefulness), both of which are derived from CS student attitudes. These factors have been found to impact retention in science and related fields [15]. For instance, [15] suggests that students persist in engineering education based on their (a) expectations of success (e.g., confidence) and (b) the perceived value and costs of an engineering degree (e.g., perceived usefulness). Other studies have also highlighted the positive effects of high confidence and understanding of the usefulness of CS student learning [8], [9], [14], [16].

B. Predicting CS Learning Performance

Predicting CS learning performance in a CS course or program creates opportunities to improve educational outcomes [17]. Through efficient performance prediction methods, educators will be able to allocate resources and instruction more accurately. Research in this area seeks various purposes, including identifying features for predictions, discovering algorithms to enhance predictions, quantifying student performance features, determining interrelated features, and identifying the underlying reasons why certain features outperform others [17]. Hellas et al. [17] conducted a systematic literature review to explore predicting student performance in CSE, revealing that the majority of the studies focused on predicting course grade (38%), individual exam grade (14.7%), program retention or dropout (13.4%), GPA or cumulative GPA (12.2%), and assignment performance (11.4%). A smaller number of articles investigated measures aiming to better quantify learning performance, such as knowledge gain or the speed at which students complete assignments.

Different types of data sources were used to extract data and predict student learning performance in the context of CSE [18]. Some studies used click-stream activities and self-reported characteristics, such as gender or learning goals [18], [19]. Hellas et al. [17] classified features used for predicting student performance into demographic (e.g., age, gender), personality (e.g., self-efficacy, self-regulation), academic (e.g., high-school performance, course performance), behavioural (e.g., log data), and institutional (e.g., high-school quality, teaching approach) factors. They noticed that the use of data describing student behaviour in a course (log data) is still relatively rare in CSE research.

Further, previous studies in CSE research applied various data analysis methods to extract insights from collected data, including statistical analysis (e.g., [5]) and ML techniques (e.g. [20]). Statistical analysis has played a fundamental role in understanding relationships within the data. However, statistical analysis remains limited and needs more effort to provide a clear answer to some questions especially when a study includes multiple complicated factors [8], [21].

Consequently, researchers recently tended to adopt ML as a means of analysis. ML (supervised and unsupervised) techniques were used to discover which significant attributes a successful learner often demonstrated in a computer course, investigate retention, and explore and analyse cognitive and non-cognitive features [20]–[24]. In the CS field in HE, research using ML with student engagement data has been applied for various purposes: to explore the impact of interventions on student behaviours in online discussion forums (e.g., [25]), to automatically identify students in need of assistance (e.g., [20]), to identify groups of students exhibiting similar performance and engagement patterns in large CS classes(e.g., [26]), to predict student outcomes from student programming data (e.g., [27]), and to model the engagement states of students working on programming exercises in an intelligent tutoring system using unsupervised ML (e.g., [28]). Previous studies also used ML as a complementary goal that can ultimately increase the theoretical understanding of the issue being researched [11] (e.g., [8]).

Despite the growing research on predicting CS learning performance and the different techniques employed to investigate it, research on the relationships between student engagement factors (BE, CE, EE, and SE), confidence, perceived usefulness, and CS learning performance has not been thoroughly explored. Further, no study has applied ML techniques to the combined self-reported engagement data and student log data to predict CS learning performance. This is especially true for novice students in the CS courses in HE who may need more support to increase their engagement and learning outcomes.

III. Research Questions

This study is a part of a larger project investigating the relationship between four dimensions of engagement, student confidence in learning CS, students' beliefs in the usefulness of learning CS, and CS learning performance, using different data analysis techniques. This paper specifically focuses on using ML prediction models. The study aims to answer the following research questions:

- **RQ1:** What is the relationship between student engagement factors, student interactions in LMS (logs) and CS learning performance?
- **RQ2:** How do student engagement factors, confidence, perceived usefulness, and log data predict CS academic performance using ML?
- **RQ3:** Which student engagement features are the most useful predictor of CS academic performance?

IV. METHODS

A. Context

The study involved two introductory core courses in the undergraduate CS program in Saudi Arabia: Introduction to Programming 1 (year 1) and Programming Languages (year 2). These two courses are fundamental requirements for all CS undergraduate students, providing the initial stepping stones in their academic journey to learn programming skills in the CS field. The Introduction to Programming 1 course is designed to introduce Object-Oriented Programming (OOP) using the Java language to students with little or no prior programming experience. Throughout the course, students are required to solve programming problems in Java through the application of basic programming principles. The course encompasses an overview of computing and introduces the fundamentals of a typical object-oriented programming language: basic data types and operators, essential object-oriented concepts, wrapper classes, console input/output, file input/output, logical expressions, control structures, classes, and methods, as well as arrays and strings. The second course is Programming Languages. The objectives of this course are to explore the fundamental concepts and design issues inherent in diverse programming languages. Additionally, the aim is to familiarize students with key programming language paradigms and their features. Both of these CS courses are delivered through a face-to-face approach.

B. Data Sources

Data was collected using three different sources: self-reported data, system log data, and CS learning performance data. The self-reported scale was used to measure multidimensional student engagement factors, confidence in learning CS, and the perceived usefulness of CS [8], [9]. It comprised 6 subscales, each capturing one variable. The first four subscales included 33 student engagement items adapted from previously validated scales in CSE [8], [9]. CE scale includes eight items about students' use of deep learning strategies and self-regulated learning in CS (e.g., "I think about different ways to solve a problem.")(Cronbach'sα= .65). BE scale involves eight items about students' involvement and investment in CS classroom activities (e.g., "I stay focused in computing 4.4. Participants 23 science class.") (Cronbach's α = .72). EE scale includes ten items about students' positive and negative reactions to and value of CS learning and activities (e.g., "I enjoy learning new things about computing science topics. ") (Cronbach's α = .87). SE scale includes seven items about the value of considering others' ideas and the quality of students' interactions in CS class (e.g., "I try to work with others who can help me in computing science") (Cronbach's α = .70). The mean of all 33 items was used to measure the overall student engagement (Cronbach's α = .90). Confidence in Learning CS scale involves five items intended to measure confidence in a student's ability to learn and to perform well on CS tasks (e.g., "I think I will do well in computer science.")(Cronbach's α = .82). The perceived usefulness scale includes five items designed to measure students' beliefs about the usefulness of learning CS (e.g., "Computer science is a worthwhile and necessary subject.")(Cronbach's α = .88). Participants were asked to rate their four types of engagement towards CS, their confidence in learning CS, and their beliefs in the usefulness of learning CS using a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Further, Log data from student behaviour on the Learning Management System (LMS), specifically the Blackboard system, included monthly logs of student access to the CS

course page on the LMS during the semester, the total hits of student interactions with the course content throughout the semester, and the total number of task submissions. In addition, to measure CS learning performance, we used Grade Point Average (GPA), a widely used metric for measuring academic performance [17], ranging from A+ (the highest grade) to D (the lowest passing grade). Each letter grade was assigned a specific numerical value from 1 to 8. GPA is determined by the final course score(ranges from 100 to 60), which was calculated by combining scores from multiple evaluation components in the course. These components included quizzes, assessments, mid-term exams, lab work, and a final exam. These components were weighted differently, with quizzes contributing 10%, assessments 5%, mid-term exams 15%, lab work 20%, and the final exam accounting for 50% of the total marks.

C. Participants

A total of 77 undergraduate students in their first year (n= 47) and second year (n= 30) agreed, consented, and completed the self-reported survey. Of these participants, 48 were male, while 29 were female. All participants were native Arabic speakers, and their ages ranged from 18 to 25.

D. Procedure

During the second semester of the 2022 – 2023 academic year, we collected data from a university in Saudi Arabia. Ethical approvals were obtained from the researchers' institution and the university involved in this study. Data was collected using three sources: an online survey (self-report), log data, and CS learning performance. We asked for cooperation from instructors teaching the targeted courses, and those who responded agreed to share the online survey link with their students. Hence, students were provided with the study information, the consent form, and the link to the online survey. It was clear to the students that their participation was voluntary and that deciding not to participate would not affect their grades and relationships with their instructors, and their data would be kept confidential. Instructors also collaborated by providing the CS academic learning performance and students' log reports extracted from the LMS.

To ensure that all students in Saudi Arabia could understand the survey questions, we translated them into Arabic, their native language. To ensure accurate translation for the survey, several stages were taken. Firstly, the survey was translated into Arabic by a native speaker who speaks both Arabic and English. This is to ensure that the translation describes the intended meaning of the original instrument in English. Once the initial translation was complete, another native Arabic speaker, who is also proficient in English, thoroughly checked the translated survey for any differences or mistakes. This review helps maintain the accuracy and reliability of the translated questions. Finally, to further confirm the quality of the translation, a bilingual expert translated the Arabic version of the survey into English and compared them to ensure both of them had the same meaning.

E. Data Analysis

We used Python for data analysis. We recorded the self-reported survey items on a numerical scale ranging from 1 to 5 and reversed negatively worded items. Data analysis involves a series of different stages, starting from the data collection and data pre-processing to the implementation of various ML models. After we collected data, we pre-processed it, and then all data, including self-reports, logs, and CS academic performance, were integrated. Subsequently, mean scales were calculated for each variable. We began the analysis by performing descriptive statics to gain insights into the distribution of values within each variable, and the Kolmogorov-Smirnov statistic test was conducted to test normality.

To address the first question, we performed a correlation analysis. To address the second and third research questions, supervised and unsupervised ML analyses were used to examine the relations between (BE, CE, EE, and SE), confidence, usefulness, and CS academic learning performance and to identify the predictive power of these factors in predicting CS academic learning performance. We used Python packages from the scikit learn library to conduct ML analysis.

The K-means clustering algorithm (unsupervised ML) was used to analyse and understand engagement patterns by classifying students based on their levels of self-reported engagement, logs, and academic performance. Then, we used classification ML algorithms(supervised ML), which can be a powerful and flexible tool for analysing self-reported data to analyse psychological traits (such as engagement factors, confidence, and perceived usefulness) [8], [11], [14], [21]. We used classification algorithms, including Random Forest (RF), Decision Tree (DT), and LightGBM (LGBM), to predict CS learning performance from student engagement data (self-reported and logs). These three algorithms were selected for their suitability to perform the prediction and exploration tasks [8], and since they belong to the tree-based family of supervised ML algorithms [8], [21]. The dataset used as input features in ML models to predict CS academic performance includes student engagement factors (33 features), confidence and perceived usefulness (10 features), and log data (7 features).

Before applying ML algorithms, we employed the 10-fold cross-validation (CV) method to train and validate the model performance. In this approach, the sample data in the training set is divided into 10 equally sized folds, with one fold used as the test set and the remaining 9 folds for training. Next, we used the feature importance technique to identify features contributing most to the prediction models. The feature importance coefficient value varies based on the algorithm used; for instance, RF can capture complex nonlinear relationships between features, which helps to give importance to sets of features that may not be as significant individually. A larger value indicates greater importance of the feature to the prediction model. In this study, we used the Mean Decrease in Gini (MDG) measure to identify the most important features. MDG quantifies the influence of each feature to

reduce impurity in decision tree-based algorithms like RF. This provides insights into which features are most influential in the model's predictive performance [29]. To evaluate the performance of ML algorithms, we applied metrics including accuracy, precision, recall, and F1-score. We applied three prediction models using the CV technique and then measured their performance separately.

V. RESULTS

A. Correlation Results

Table I provides a brief overview of the descriptive statistics for the study variables.

We conducted a Spearman correlation analysis to explore the relationship between all variables (see table II). The correlation analysis results provide a comprehensive understanding of relationships among the variables in our data, ranging from strong positive correlations to weak negative correlations. Results show that all four types of student engagement have a low to moderate positive correlation with each other. Specifically, starting with CE, it shows moderate positive correlations with BE, EE, and SE ($r_s = 0.53$, $r_s = 0.63$, $r_s = 0.35$, $p < 0.001$). It also shows a weak positive correlation with usefulness ($r_s = 0.12$, $p < 0.001$). However, CE has a weak negative correlation with confidence ($r_s = -0.14$, $p < 0.001$). Moving on to BE, it has a moderate positive correlation with CE, EE, and SE ($r_s = 0.53$, $r_s = 0.69$, $r_s = 0.47$, $p < 0.001$). It also shows weak positive correlations with confidence, usefulness, and CS academic performance ($r_s = 0.13$, $r_s = 0.26$, $r_s = 0.10$, $p < 0.001$). Similarly, EE shows a moderate positive correlation with CE, BE, and SE. EE also shows weak positive correlations with usefulness ($r_s = 0.21$, $p < 0.001$). SE has a moderate positive correlation with CE, BE, and EE. Its weak positive correlations are observed with usefulness ($r_s = 0.17$, $p < 0.001$). However, SE has a weak negative correlation with CS academic performance ($r_s = -0.27$, $p < 0.001$).

The confidence factor has moderate positive correlations with the usefulness factor ($r_s = 0.43$, $p < 0.001$), and weak positive correlations with BE ($r_s = 0.13$, $p < 0.001$), and CS academic performance ($r_s = 0.12$, $p < 0.001$). The usefulness has a moderate positive correlation with confidence. It also shows a weak positive correlation with CE, BE, EE, SE, and overall engagement($r_s = 0.12$,$r_s = 0.26$,$r_s = 0.21$,$r_s = 0.17$,$r_s = 0.23$, $p < 0.001$). However, it has negative correlations with CS academic performance. Student Log factor has a weak positive correlation with CS academic performance ($r_s = 0.10$, $p < 0.001$). CS academic performance has a weak positive correlation with BE, Logs, and confidence ($r_s = 0.10$, $r_s = 0.12$, $r_s = 0.10$, $p < 0.001$). However, It has a weak negative correlation with SE and usefulness.

B. Clustering Results

We used the K-means clustering algorithm to investigate patterns of engagement among students by classifying them into different groups based on their self-reported levels of engagement, logs, and CS academic performance. By grouping students into clusters, we were able to investigate different engagement factors and how they relate to CS academic outcomes.

To determine the optimal number of clusters (K), we utilized both the elbow method and silhouette score. The elbow method helps in selecting the appropriate value of K by identifying the point where the within-cluster sum of squares begins to stabilize. In our analysis, the elbow method suggested that the optimal number of clusters is 6. Additionally, we utilized silhouette scores to further validate this selection. The silhouette score, which measures the cohesion and separation of clusters, yielded a value of 0.54, indicating a reasonable level of clustering effectiveness. Subsequently, we applied K-means clustering using the selected value of K (k=6) to group the students accordingly.

Table III represents the K-Means clustering results showing the characteristics of each cluster. It is clear from the results presented in table III that the group of students who show active engagement with the CS course content(high logs) tends to achieve higher grades.

Specifically, students in cluster 1 show high engagement across behavioural, emotional, and social engagement dimensions, high confidence, and find the course very useful, the highest level of activity (logs), and the highest grades among all clusters. Students in cluster 0 show high levels of engagement across behavioural and emotional engagement, high confidence and perceived usefulness, low levels of activity (logs), and achieve moderate grades. Students' engagement levels in cluster 2 are moderate across different dimensions, with high confidence and perceived usefulness, moderate activity (logs), and moderate grades. Students' engagement levels in cluster 3 are high, with behavioural engagement being the highest, high confidence and perceived usefulness, low activity (logs), and moderate grades. Students in cluster 4 show high behavioural and social engagement levels, moderate confidence and usefulness, low activity (logs), and low grades. Student engagement levels in cluster 5 are moderate across all dimensions except emotional engagement, moderate confidence, and high perceived usefulness. However, this cluster has the lowest activity (logs) and lowest grades among all clusters.

C. Classification Results

Classification algorithms using RF, DT, and LGBM were used to predict CS learning performance. This study's dataset had 49 features(self-reported and logs). We reported the performance measures for each classifier. Table IV presents accuracy, precision, recall, and F1-score for each ML classifier.

Among these models, RF and LightGBM achieved the highest accuracy scores, both at 0.45, indicating that they correctly predicted CS learning performance with 45% accuracy. When considering precision, recall, and F1-score metrics, RF outperformed the other models, with higher precision and F1-scores at 0.42 and 0.43, respectively.

MDG measure within the RF model was used to find the most important features. Figure 1 presents features as ranked

TABLE I
DESCRIPTIVE STATISTICS

Variables	**Mean**	**Std**	**Min**	**Max**
CE	3.50	0.46	2.50	4.50
BE	3.97	0.56	2.50	5.00
EE	4.03	0.80	2.30	5.00
SE	3.94	0.73	2.14	5.00
Confidence	4.03	0.73	1.80	5.00
Usefulness	4.59	0.52	2.80	5.00
Logs	43.13	22.19	9.50	99.83
Overall Engagement	3.86	0.50	2.99	4.84
CS learning performance	75.32(3.92)	12.44(2.42)	60.00(1.00)	99.50(8.00)

TABLE II
SPEARMAN CORRELATION

Variables	1	2	3	4	5	6	7	8	9
CE	1.00	0.53	0.63	0.35	-0.14	0.12	0.05	0.73	0.03
BE	-	1.00	0.69	0.47	0.13	0.26	-0.05	0.82	0.10
EE	-	-	1.00	0.53	-0.07	0.21	-0.02	0.89	0.07
SE	-	-	-	1.00	-0.05	0.17	-0.02	0.77	-0.27
Confidence	-	-	-	-	1.00	0.43	0.08	-0.05	0.12
Usefulness	-	-	-	-	-	1.00	0.06	0.23	-0.14
Logs	-	-	-	-	-	-	1.00	-0.02	0.10
Overall Engagement	-	-	-	-	-	-	-	1.00	-0.05
CS academic performance	-	-	-	-	-	-	-	-	1.00

** Correlation is significant at the 0.01 level (2 - tailed).

TABLE III
CLUSTER STATISTICS

Cluster	**CE**	**BE**	**EE**	**SE**	**Confidence**	**Usefulness**	**Logs**	**Grade**
0	3.55	4.00	4.08	3.79	4.12	4.62	36.86	4.37
1	3.47	4.00	4.48	4.21	4.15	4.25	98.63	5.00
2	3.40	3.80	3.77	3.73	4.07	4.60	72.48	4.55
3	3.40	4.03	4.07	4.01	4.10	4.39	25.61	4.00
4	3.65	4.00	3.97	4.13	3.84	4.76	51.50	3.19
5	3.46	3.95	4.06	3.92	3.91	4.86	13.26	2.57

TABLE IV
PERFORMANCE METRICS OF CLASSIFICATION ML ALGORITHMS

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-score**
DT	0.40	0.38	0.38	0.39
RF	0.45	0.42	0.45	0.43
LGBM	0.45	0.44	0.45	0.44

by MDG within RF to predict CS learning performance. The top 6 important features were: 'Total-H'(total hits of student interactions within the content of the course in LMS throughout an academic semester),' Total Access3'(the frequency of students accessing the LMS during the third month of the semester), 'Total Access4'(the frequency of students accessing the LMS during the fourth(last) month of the semester),' Total Access1'(the frequency of students accessing the LMS during the first month of the semester),' Total Access2' (the frequency of students accessing the LMS during the second month of the semester), and the total of submissions. These log features likely have a strong relationship with the CS learning performance variable and could provide valuable insights for understanding the underlying features that drive the prediction.

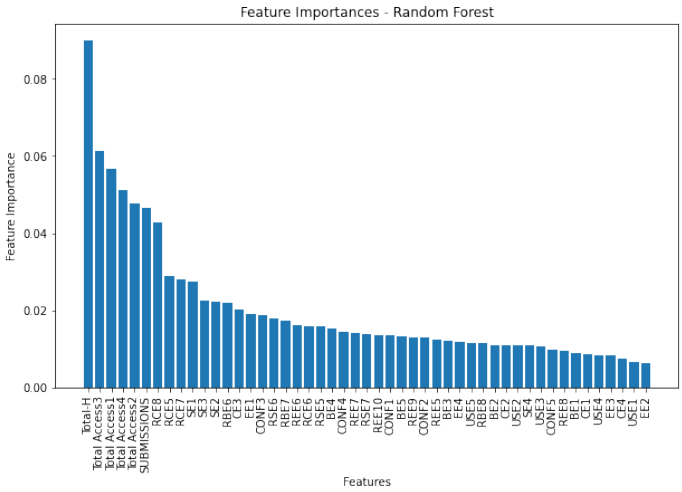

Fig. 1. Important features

VI. DISCUSSION

The research provides new insights into understanding the multi-dimensions of student engagement, students' confidence in learning CS, their perceived usefulness of CS, and their effects on CS academic performance. This study is the first to examine self-reported student engagement factors, confidence, perceived usefulness, and log data and analyse them using ML techniques, with a particular focus on assessing the predictive power of these factors on CS academic performance.

A. What is the relationship between student engagement factors, students' confidence in learning CS, perceived usefulness of CS, student interactions in LMS (logs), and CS learning performance?

The results of the study show complex correlations between different dimensions of student engagement, confidence, perceived usefulness, and academic performance in CS courses. Particularly, CE has moderate positive correlations with BE, EE, and SE, indicating that students who actively participate, are emotionally engaged, and engage socially tend to be more cognitively involved in the course. These results highlight the interrelation of different dimensions of engagement within the CS learning environment [9]. These findings align with literature emphasising the multi-dimensional nature of student engagement and its impact on academic learning outcomes [7].

In terms of students' confidence in their programming skills and perceived usefulness, confidence appears as a significant factor positively associated with perceived usefulness and, to a less significant extent, with BE and CS academic performance. This suggests that students who have confidence in their abilities are more likely to perceive the course as beneficial and may perform better academically [8], [9], [14], [16]. However, the perceived usefulness shows a weak negative correlation with CS academic performance, indicating that while students may find the course useful, this perception does not necessarily tend to improve academic performance [30]. Further, the negative correlation found between SE and perceived usefulness, alongside academic performance, suggests that both social engagement and perceived usefulness are important factors that may influence other aspects affecting students' success in CSE. These findings call for further investigation to understand the relationship between perceived usefulness and academic achievement in CSE [31].

The findings also reveal a significant positive correlation between student logs, student interactions with course materials in LMS, and CS academic performance. This result highlights the important role of active learning and engagement with course content in achieving better academic outcomes [5]. This result is consistent with existing literature on student engagement, highlighting the importance of active participation and interaction with learning materials in enhancing academic achievement [5], [32].

B. How do student engagement factors, confidence, perceived usefulness, and log data predict CS academic performance using ML?

Student engagement factors (CE, BE, EE, and SE), confidence, perceived usefulness, and log data were analysed using supervised and unsupervised ML algorithms to predict CS learning performance. Using K-Means unsupervised ML, we were able to investigate different patterns of engagement among students by classifying them into six different groups based on their self-reported levels of engagement, confidence, perceived usefulness, logs, and CS academic performance. This approach is consistent with prior research that highlights the multidimensional nature of student engagement [8], [9] and its impact on academic outcomes [7]. Clustering results highlight that a group of students who show active engagement with the CS course content (high logs) tends to achieve higher grades. This aligns with literature suggesting that active engagement, as evidenced by frequent interactions with course content, often correlates with enhanced academic achievement [33], [34]. On the other hand, another group with a lower engagement with the learning platform (lower access log count) had lower academic performance outcomes. These results align with prior studies highlighting the significance of engagement behaviours, such as logging activity, as predictors of academic success in online learning environments [33], [34]. Moreover, our study highlights the advantage of clustering techniques in discovering engagement factors among students, providing valuable insights for educators and policymakers to design interventions and support strategies effectively [19].

Classification using supervised ML algorithms using RF, DT, and LGBM were used to predict CS learning performance from student engagement data (self- reported and logs).RF outperformed the other models, with 45% accuracy, precision(0.42), and F1-scores (0.43). These performance scores indicate that the RF model could predict the CS learning performance more accurately than the other tested models. This finding aligns with previous research in the educational domain, demonstrating the superior performance of the RF model over other ML prediction models, likely due to their adeptness in managing high-dimensional and non-linear relationships among student engagement features [8], [10], [35], [36]. However, it is worth noting that all models achieved relatively moderate accuracy scores, suggesting that other factors beyond the features considered in the models might influence CS learning performance. Further exploration and refinement of features or the application of more advanced ML modelling techniques could potentially improve prediction accuracy and provide deeper insights into the engagement factors predicting CS learning outcomes.

C. Which student engagement features are the most useful predictor of CS academic performance?

Study results highlight the key student engagement indicators that predict CS academic performance, as identified through the use of the RF classification prediction model. The

results revealed that log data were more successful in predicting students' academic performance than self-reported data. This result aligns with the findings of [19], suggesting that features of self-reported data were less effective in predicting students' academic achievement if compared to the features generated from Blackboard log data and engagement activities.

Several key features that were extracted from student log data within CS course content in the Blackboard system emerged as significant predictors of CS learning performance. These log features likely have a strong relationship with the CS learning performance variable and could provide valuable insights for understanding the underlying features that drive the prediction. Specifically, the study results highlighted various log features associated with academic achievement in CS, including the total number of student interactions within course content in the LMS throughout the academic semester, as well as the frequency of students accessing the LMS during different stages of the semester. Accessing the course content during the initial two months of the semester emerged as an influential factor, indicating early engagement patterns that may set the trajectory for successive academic outcomes [27], while their interaction in the third and fourth months suggests the importance of sustained engagement over the duration of the course. Moreover, the total number of submissions, reflecting students' active completion of course assignments and other tasks, emerged as an important predictor of CS learning performance. These findings align with previous research emphasising the significance of student engagement behaviours in shaping academic success [2].

VII. IMPLICATIONS

These findings are important for CS educators, institutions, researchers, and any stakeholders who want to improve engagement in CSE, as examining novice student engagement and understanding students' attitudes toward CS in the early stage of their studies may improve retention of students in the CS field. The study findings imply that ML models can be valuable tools for analysing and predicting student learning outcomes based on various engagement factors. The use of a multi-dimensional engagement instrument and the ML techniques employed may guide other researchers in developing interventions to improve CS learning outcomes and further our understanding of the phenomenon. Furthermore, CS educators could use the identified engagement features to improve their students' engagement and learning outcomes in CS classes. For instance, educators could potentially focus on encouraging student interactions with the course content(BE) if they want to increase student engagement in learning CS, especially for novice students [5]. This could involve creating a flexible learning environment where students are encouraged to participate, ask questions, provide help to others, and discuss CS topics outside of class. Moreover, CS educators can use effective learning strategies within LMS to create dynamic and engaging learning experiences. For example, they can integrate pair programming exercises directly into the LMS, allowing students to collaborate in solving CS problems. Additionally, educators can use peer instruction, group problem-solving activities, and flexible quiz activities within the LMS to improve students' engagement and learning experiences in CS courses [37].

VIII. LIMITATIONS AND FUTURE WORK

While this work has provided interesting insights into the topics under investigation, our research had some limitations. Firstly, the models built in this study did not perform very well. One possible reason is that engagement factors selected in this study may not have been enough to predict the CS learning performance due to the limited nature of the data source (self-report instruments and limited log data). Further, one of the important limitations of this study is that some log data features, such as time attributes (like the amount of time spent submitting assignments using the Blackboard system), were not included in the analysis. This absence occurred as the researchers lacked access to the system, and these features were not present in the provided log files. Future work should consider strategies to include more log data from LMS to complement the measure of student engagement [38]. However, such data may pose some challenges in terms of ethical considerations, as access to students' log data with LMS requires more stringent ethical procedures [39]. The second reason could be the sample size. Future research should consider larger data sets from different sources and different institutions. Future studies should employ additional machine learning algorithms, exploring advanced ensemble methods and hyperparameter tuning, as well as deep learning techniques to enhance predictive performance.

IX. CONCLUSION

This study used ML algorithms to analyse student engagement in CS and predict the CS learning outcomes of novice students in CS classes. This is one of the first studies to examine this area among novice students in Saudi Arabia in CSE. Our results indicate that there is a strong positive relationship between behavioural engagement, confidence logs factors, and CS learning performance. Further, it was possible to predict CS learning performance from self-reported levels of engagement, logs, and CS academic performance. We found that ML models, particularly RF, were useful in making predictions and extracting valuable information. The important features that impacted prediction were extracted from log data. The implications of this study are significant for educational professionals and policymakers interested in promoting novice student engagement and learning outcomes in CSE. Further research is needed to explore the generalisability of these research findings.

REFERENCES

[1] H. Yildiz Durak, "The effects of using different tools in programming teaching of secondary school students on engagement, computational thinking and reflective thinking skills for problem solving," *Technology, Knowledge and Learning*, vol. 25, pp. 179–195, 2020.

[2] S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, and L. Porter, "A robust machine learning technique to predict low-performing students," *ACM transactions on computing education (TOCE)*, vol. 19, no. 3, pp. 1–19, 2019.

[3] M. Morgan, M. Butler, J. Sinclair, C. Gonsalvez, and N. Thota, "Contrasting cs student and academic perspectives and experiences of student engagement," in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 2018, pp. 1–35.

[4] M. Morgan, J. Sinclair, M. Butler, N. Thota, J. Fraser, G. Cross, and J. Jackova, "Understanding international benchmarks on student engagement: awareness and research alignment from a computer science perspective," in *Proceedings of the 2017 ITiCSE Conference on Working Group Reports*, 2018, pp. 1–24.

[5] A. S. Carter, C. D. Hundhausen, and O. Adesope, "Blending measures of programming and social behavior into predictive models of student achievement in early computing courses," *ACM Transactions on Computing Education (TOCE)*, vol. 17, no. 3, pp. 1–20, 2017.

[6] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of educational research*, vol. 74, no. 1, pp. 59–109, 2004.

[7] M.-T. Wang, J. A. Fredricks, F. Ye, T. L. Hofkens, and J. S. Linn, "The math and science engagement scales: Scale development, validation, and psychometric properties," *Learning and Instruction*, vol. 43, pp. 16–26, 2016.

[8] S. A. A. Albakri, M. B. Ada, and A. Morrison, "Exploring student engagement, confidence, and usefulness for female students in cs class at high school using machine learning," in *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2023, pp. 1–9.

[9] S. Albakri, M. B. Ada, and A. Morrison, "The roles of confidence and perceived usefulness in female student engagement in high school computing science," in *Proceedings of the 18th WiPSCE Conference on Primary and Secondary Computing Education Research*, 2023, pp. 1–9.

[10] S. Joel, P. W. Eastwick, C. J. Allison, X. B. Arriaga, Z. G. Baker, E. Bar-Kalifa, S. Bergeron, G. E. Birnbaum, R. L. Brock, C. C. Brumbaugh *et al.*, "Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies," *Proceedings of the National Academy of Sciences*, vol. 117, no. 32, pp. 19 061–19 071, 2020.

[11] T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning," *Perspectives on Psychological Science*, vol. 12, no. 6, pp. 1100–1122, 2017.

[12] S. A. A. Albakri, "Using machine learning algorithms for analysing the factors that affect pupil engagement and learning outcomes in cse," in *Proceedings of the 2022 Conference on United Kingdom & Ireland Computing Education Research*, 2022, pp. 1–1.

[13] J. P. Connell, "Context, self, and action: A motivational analysis of self-system processes across the life span." 1990.

[14] A. Hoegh and B. M. Moskal, "Examining science and engineering students' attitudes toward computer science," in *2009 39th IEEE frontiers in education conference*. IEEE, 2009, pp. 1–6.

[15] E. Seymour and N. M. Hewitt, *Talking about leaving*. Westview Press, Boulder, CO, 1997, vol. 34.

[16] Y. J. Xu, "Career outcomes of stem and non-stem college graduates: Persistence in majored-field and influential factors in career choices," *Research in Higher Education*, vol. 54, pp. 349–382, 2013.

[17] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: a systematic literature review," in *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, 2018, pp. 175–199.

[18] E. Vinker and A. Rubinstein, "Mining code submissions to elucidate disengagement in a computer science mooc," in *LAK22: 12th international learning analytics and knowledge conference*, 2022, pp. 142–151.

[19] A. D. Ali and W. K. Hanna, "Predicting students' achievement in a hybrid environment through self-regulated learning, log data, and course engagement: A data mining approach," *Journal of Educational Computing Research*, vol. 60, no. 4, pp. 960–985, 2022.

[20] A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen, "Exploring machine learning methods to automatically identify students in need of assistance," in *Proceedings of the eleventh annual international conference on international computing education research*, 2015, pp. 121–130.

[21] C.-Y. Ko and F.-Y. Leu, "Examining successful attributes for undergraduate students by applying machine learning techniques," *IEEE Transactions on Education*, vol. 64, no. 1, pp. 50–57, 2020.

[22] N. Gao, W. Shao, M. S. Rahaman, and F. D. Salim, "n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–26, 2020.

[23] X. Wang, A. Dourado, P. B. Thomas, and C. R. Bego, "Modeling engineering persistence through expectancy value theory and machine learning techniques," in *2022 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2022, pp. 1–9.

[24] C. Herbert, A. El Bolock, and S. Abdennadher, "How do you feel during the covid-19 pandemic? a survey using psychological and linguistic self-report measures, and machine learning to investigate mental health, subjective experience, personality, and behaviour during the covid-19 pandemic among university students," *BMC psychology*, vol. 9, no. 1, pp. 1–23, 2021.

[25] D. La Vista, N. Falkner, and C. Szabo, "Understanding the effects of intervention on computer science student behaviour in online forums," *City*, 2017.

[26] H. Khosravi and K. M. Cooper, "Using learning analytics to investigate patterns of performance and engagement in large classes," in *Proceedings of the 2017 acm sigcse technical symposium on computer science education*, 2017, pp. 309–314.

[27] O. H. Lu, J. C. Huang, A. Y. Huang, and S. J. Yang, "Applying learning analytics for improving students engagement and learning outcomes in an moocs enabled collaborative programming course," in *Learning analytics*. Routledge, 2018, pp. 78–92.

[28] H. L. Fwa and L. Marshall, "Modeling engagement of programming students using unsupervised machine learning technique," *GSTF Journal on Computing*, vol. 6, no. 1, p. 1, 2018.

[29] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[30] J. L. Burnette, C. L. Hoyt, V. M. Russell, B. Lawson, C. S. Dweck, and E. Finkel, "A growth mind-set intervention improves interest but not academic performance in the field of computer science," *Social Psychological and Personality Science*, vol. 11, no. 1, pp. 107–116, 2020.

[31] M. Barr and M. Kallia, "Why students drop computing science: Using models of motivation to understand student attrition and retention," in *Proceedings of the 22nd Koli Calling International Conference on Computing Education Research*, 2022, pp. 1–6.

[32] B. Hoffman, R. Morelli, and J. Rosato, "Student engagement is key to broadening participation in cs," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019, pp. 1123–1129.

[33] D. Azcona and A. F. Smeaton, "Targeting at-risk students using engagement and effort predictors in an introductory computer programming course," in *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings 12*. Springer, 2017, pp. 361–366.

[34] D. Azcona, I.-H. Hsiao, and A. F. Smeaton, "Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints," *User Modeling and User-Adapted Interaction*, vol. 29, pp. 759–788, 2019.

[35] 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.

[36] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced mooc course using random forest model," *Information*, vol. 12, no. 11, p. 476, 2021.

[37] M. L. Maher, C. Latulipe, H. Lipford, and A. Rorrer, "Flipped classroom strategies for cs education," in *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, 2015, pp. 218–223.

[38] C. R. Henrie, L. R. Halverson, and C. R. Graham, "Measuring student engagement in technology-mediated learning: A review," *Computers & Education*, vol. 90, pp. 36–53, 2015.

[39] L. A. Macarini, C. Cechinel, H. L. d. Santos, X. Ochoa, V. Rodés, G. E. Alonso, A. P. Casas, and P. Díaz, "Challenges on implementing learning analytics over countrywide k-12 data," in *Proceedings of the 9th international conference on learning analytics & knowledge*, 2019, pp. 441–445.