IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 14, 2021, accepted September 27, 2021, date of publication October 1, 2021, date of current version October 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117117

Deep Learning Model to Predict Students Retention Using BLSTM and CRF

DIAA ULIYAN[1], ABDULAZIZ SALAMAH ALJALOUD[1], ADEL ALKHALIL[1], HANAN SALEM AL AMER[2], MAGDY ABD ELRHMAN ABDALLAH MOHAMED[3,4], AND AZIZAH FHAD MOHAMMED ALOGALI[5,6]

[1]Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia
[2]Department of Curriculum and Teaching Methods, College of Sciences, University of Ha'il, Ha'il 81481, Saudi Arabia
[3]Foundations of Education Department, Community College, University of Ha'il, Ha'il 81481, Saudi Arabia
[4]Education College, New Valley University, Kharga Oasis 72511, Egypt
[5]Department of Educational Leadership, University of Rochester, Rochester, NY 14627, USA
[6]Department of Educational Leadership, The University of Akron, Akron, OH 44325, USA

Corresponding author: Diaa Uliyan (d.uliyan@uoh.edu.sa)

The authors would like to acknowledge the support and funding of this project by the Deanship of Scientific Research at the University of Ha'il, Saudi Arabia, under Grant ''RG-20 146''.

ABSTRACT There is an increasing awareness that predictive analytics helps universities to evaluate students' performances. Big data analytics, such as student demographic datasets, can provide insight that helps to support academic success and completion rates. For example, learning analytics is an essential component of big data in universities that can provide strategic decision makers with the opportunity to perform a time series analysis of learning activities. A two-year retrospective analysis of student learning data from the University of Ha'il was conducted for this study. Predictive deep learning techniques, the bidirectional long short term model (BLSTM), were utilized to investigate students whose retention was at risk. The model has diverse features which can be utilized to assess how new students will perform and thus contributes to early prediction of student retention and dropout. Further, the condition random field (CRF) method for sequence labeling was used to predict each student label independently. Experimental results obtained with the predictive model indicates that prediction of student retention is possible with a high level of accuracy using BLSTM and CRF deep learning techniques.

INDEX TERMS Student retention, data analytics, bidirectional long short term, condition random field, deep learning.

I. INTRODUCTION

An enduring challenge in higher education all around the world is student retention [1]. Simply, higher education institutions are increasingly aware of the critical need to develop innovative approaches that ensure students graduate in a timely fashion and are well trained and workforce-ready in their field of study. As the volume and variety of data collected in both traditional and online university offerings continues to expand, new opportunities arise to apply big data analytics to challenges in higher education. Student retention is most commonly conceptualized as year-by-year retention or persistence rates as well as graduation rates [2]. Together, these rates indicate student success rates, which are typically defined as primary key indicators of university performance. In addition, they reflect the overall quality of student learning behaviour.

Prior research has shown that the decision by a student to voluntarily withdraw from their course of study may be influenced by both personal and institution-related factors [3], [4]. Regarding institutional factors specifically, higher education institutions increasingly recognize that student retention rates can be important indicators of student satisfaction with the institution and/or the learning curriculum [5], [6]. The present study examined the relationship between the preparatory year (bridging year) program at the University of Ha'il, undertaken by students prior to commencing their bachelor's degree, and the retention rate of these students. As such, its findings contribute to a deeper level of understanding of the role of institution-led pre-degree academic preparatory programs in the graduation outcomes of students. It also provides a unique practical contribution to the strategic planning relevant to our

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

understanding of the relationship between the decisions made by students to remain enrolled on their course of study and the main factors that influence their decision making. Three research questions were addressed in this study: 1) What are the main features of students at the early stages of study which help to indicate student retention rates?; 2) How can student retention rates be improved?; and 3) What is the impact of first-year course grades on graduation rates for undergraduate students in a given period?

This paper focuses on two key performance indicators that are usually used in universities, that are particularly important to any investigation of student behaviours because they indicate whether they may be at risk of discontinuing their studies:

i. *First-year student retention rates* - defined as the percentage of first-year undergraduate students who continue at the university through to the next year in relation to the total number of first-year students in the same year.
ii. *Graduation rate for undergraduate students in a given period*- defined as the percentage of undergraduate students in each subject cohort who completed the programs during the specified period (no more than one successive year out of university).

The main objectives of this paper are as follows: 1) to identify students at risk of failure to assist universities to devise an early intervention plan to amend student performance correctly and prevent student drop out; and 2) to ascertain the effectiveness of the deep the bidirectional long short term model (BLSTM) model with the condition random field (CRF) methods, for the early prediction of students at completion risk, compared to conventional approaches.

II. LITERATURE REVIEW

The common elements, related to student retention, are defined as follows:

A. STUDENT RETENTION AND ITS RELATIONSHIP TO STUDENT PERFORMANCE

Student retention refers to the outcome whereby, once enrolled in their course, students ''remain and successfully complete their studies'' [5]. The literature typically measures student graduation in terms of retention rates over a period of 5-6 years from the time of enrolment [7]. There is general agreement within the field of higher education that ''student retention arises from a complex combination of student, institutional and external factors'' [8]. These factors manifest in different ways for different students and, as a result, there is a complex relationship between the contextual factors that underpin a student's decision to stay enrolled in or to withdraw from a course and the performance of the student. Notably, student retention in higher education is invariably linked in the literature to student performance while completing their studies as well as to the university enrolment system and the acceptance of students on particular courses [9].

Findings from research conducted in the Saudi higher education context show that student retention may be influenced by personal, social, academic, and institutional factors [4], [10]. Personal and social factors refer to those aspects related to the students' personalities and their external communities, respectively and may include such things as learning motivation, ability to manage the demands of the course, family support, financial status, and the like [10]. Academic factors relate to the learning program and materials the students engage with during the course of study, whereas institutional factors are related to the policies and procedures of the institutions such as admission policies and the provision of student support services [10]. In addition, research shows that factors such as student ethnicity, race, and gender, as well as the distance from the student's hometown to the learning institution, can be important predictors of student retention outcomes [3], [4].

B. GRADUATE RETENTION

Included in the broader construct of student retention is graduate retention. This refers to students who have graduated from their undergraduate degree and choose to pursue postgraduate studies. The graduate retention rate therefore refers to the number of graduated students who are retained throughout their postgraduate study program [11].

C. STUDENT ENGAGEMENT

In the field of education, student engagement broadly refers to the level of attention, interest, and motivation the student displays towards the learning materials or topic [2]. As such, it is a complex and multidimensional construct that is strongly correlated to student retention [8]. As a generalization, students who demonstrate positive engagement with the learning materials and/or topic are more likely to achieve successful learning outcomes and to complete their courses of study [2]. Similar to student retention, student engagement is potentially influenced by multiple factors and may therefore be experienced in different ways by different students across different learning contexts [8].

D. STUDENT ATTRITION/STUDENT DROPOUT

Terms relevant to a lack of student retention include student withdrawal, student attrition, and student dropout. These terms refer to the circumstance where the student has been unable to complete his/her studies and has subsequently left the course prior to its completion [12].

The complex interrelationship of factors influencing student retention in higher education is also increasingly influenced by advancements in both information and communication technologies and the capabilities of universities to collect, manage, and analyze student enrolment data. As [8] suggests, ''increased network capabilities, machine learning and artificial intelligence are poised to fundamentally impact on the relationship between students, teachers and institutions.''

E. IMPORTANCE OF STUDENT RETENTION

The broader literature generally acknowledges that student retention rates are a major consideration for higher education institutions worldwide [5], [6], [8]. Although it also acknowledged that students may choose to withdraw voluntarily from their study program, there may also be institutional factors or elements of the curriculum design that prompt a withdrawal [3], [10]. As a result, tertiary institutions around the world realize the need to have in place strategies and plans to monitor and address student attrition factors [6]. Moreover, tertiary institutions around the world are aware of the increasing student dropout rates. For instance, the American College Testing (ACT) Report on student retention and graduation rates in colleges in the United States (US) from 1991 to 2012 reported that student graduations (within 5 years of course commencement) across all college institutions were at 51.9% in 2012, a decrease from 54.4% in 1991 [13].

In terms of the evidence around the causes of student retention, research shows that student experiences of the culture of the university and the pedagogical approaches implemented by teachers can have a significant influence. For instance, a study by [14] of 265 university students attending colleges in the US found that students who did not return the following year for study or who changed their major to another field had significantly lower perceptions of social connectedness and satisfaction with faculty members (approachability and quality of interactions) compared to students who returned. Further to the relationship between student-faculty member interactions and student retention, [15] shows that the number and quality of meetings held with an academic advisor can also impact first-year university students' decisions to stay enrolled in the course. Indeed, [15] concluded from the study of 363 first-year students studying at universities in the US from Fall 2009 to Fall 2010 that every satisfying meeting with an academic advisor increases the odds that a student will remain in the course by 13%.

A recent study by [16] focused on the influence of financial stress, debt levels, and the availability of financial counselling on the retention rates of 2,475 undergraduate students studying in the US. The researchers found that both financial stress and student loan debt contributed to an increased likelihood of withdrawal from college. Moreover, they found that students who had sought out financial counselling were more likely to withdraw from college within the following year compared to students who had not accessed financial counselling [15].

Given the ongoing advancements in technologies for collecting, storing, managing, and mining data it is not surprising that universities are increasingly utilizing them to be better informed about student retention outcomes. A study by [17], for instance, applied three data mining methods: artificial neural networks, decision trees, and logistic regression, to 8 years' worth of institutional data from a university in the US mid-west region to predict the retention/attrition rates of 25,224 freshmen students. The researcher reported that the artificial neural networks performed the best (81% prediction accuracy) and found that educational and financial variables were the most important predictors of student attrition.

Moreover, for universities, an important outcome to emerge from advancements in technologies for teaching and learning is the growth of online learning platforms. Online learning provision has potentially important implications for student retention rates at universities. For instance, it has potential as a learning pathway to overcome some of the factors influencing a student's decision to withdraw from their studies such as distance to university [18]. Conversely, it has the potential to accentuate some factors leading to student withdrawal such as the feeling of not having access to adequate support [19]. In terms of the research evidence, a study recently conducted by [20] compared university students' ($n = 15$) and university faculty members' ($n = 15$) perceptions of the main factors that influenced student retention in online education. The researchers reported that the top five factors, according to the students, were increased faculty instruction, the provision of meaningful feedback, accessing course credits for previous study, achieving a desired grade point average (GPA), and the provision of institutional support. The top five factors to influence student retention on online courses, according to the faculty members, were student self-discipline, faculty-student interaction quality, the provision of institutional support, grade received, and accessing transfer credits [20].

In response to the complex interrelationship of factors influencing student retention rates in higher education, institutions across the sector have an interest in better understanding the types of infrastructure and preparatory programs required to improve student retention. Reference [21] conducted a study to identity the characteristics of students most associated with community college retention. In addition, they examined the relationship between student participation in a preparatory study skills course and overall retention rates. The study sample comprised 1,740 freshman students from a community college with three campuses across four county districts in the US. In terms of student characteristics, the researchers found gender followed by age were the most significant predictors of retention. Specifically, females were more likely to be retained than males and students aged 40 years and above were more likely to be retained than those aged 18-39. Regarding the retention prediction significance related to participation in a study skills course, the researchers reported that participation was a significant predictor of retention. That is, the student participants who successfully completed the study skills course were 63.6% more likely to be retained compared to students who did not take the course.

III. PROPOSED METHOD

Improvements to retention rates can be initiated by creating a student demographic dataset as a measure of undergraduate student performance at the university. This data consists of

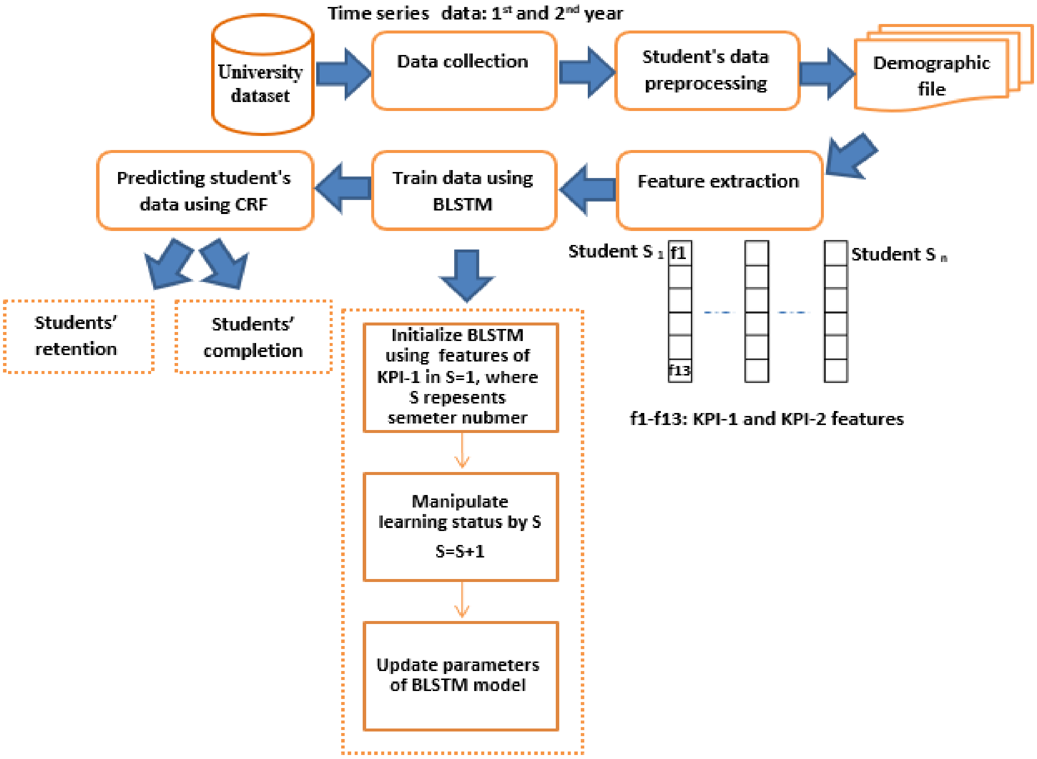

Time series data: 1st and 2nd year
University dataset
Data collection
Student's data preprocessing
Demographic file
Predicting student's data using CRF
Train data using BLSTM
Feature extraction
Student S 1
f1
f13
Student S n
Students' retention
Students' completion
Initialize BLSTM using features of KPI-1 in S=1, where S repesents semeter nubmer
Manipulate learning status by S
S=S+1
Update parameters of BLSTM model
f1-f13: KPI-1 and KPI-2 features

FIGURE 1. Graphical representation of the proposed framework.

university related information collected on the students. This dataset typically includes student grade averages, standardized assessment test results, participation rates, and attendance. It can also be used to measure student perceptions of the university and to predict retention. It is useful for the university to collect and store student demographic data for statistical analysis, such as machine learning, to predict student retention as a key indicator in the university quality assurance process [22]. This paper examines two performance indicators: first-year student completion rates and student graduation rate.

The prediction method in this research can be summarised as follows:

1. Collect student data from the university dataset during the first and second years of their study.
2. Adopt a sampling method to preprocess a subset of the student data'. Or do you mean, 'Adopt a sampling method to select a subset of the student data. The sampling criterion used was "college = computer science and engineering". Save the data in a demographic file for use in the prediction process.
3. Extract relevant features from student's data and save them in feature vectors: KPI-1 and KPI-2 features.
4. Train features using a bidirectional LSTM learning network to determine optimal weights.
5. Label the output of BLSTM features using CRF to predict each student's label independently.

The proposed method is illustrated in **Fig. 1**.

A. DATA COLLECTION AND PREPROCESSING

Student engagement in the university environment is a major concern for higher education institutions due to its implications for student retention. A better understanding of this issue can be achieved by selecting data from the university dataset using structured query language (SQL). Then a sampling approach is employed to select only significant features of students during their first and second years. For instance, in the first year, information stored about students includes their preparatory GPA, Math 1 Gr, Eng 1 Gr and assessment outcomes from the first year. These features are saved in a student demographic data file. The demographic data held for each student could be increased to include other important features such as second year course grades, GPAs of semesters 3 and 4, etc. The key reason for using demographic data is to perform valuable and effective selection from a large amount of data without noise. Hence, we need to extract features from the university's back-end dataset and transform the retention prediction problem into a time series prediction problem.

B. FEATURE EXTRACTION

The main aim of preprocessing student information is to remove incorrect data from the dataset [23]. This step is mandatory for extracting relevant features from the data. The features are described in Table 1. Two dimensions of student features have been extracted as key performance indicators (KPIs) in this paper: graduation rate and first year

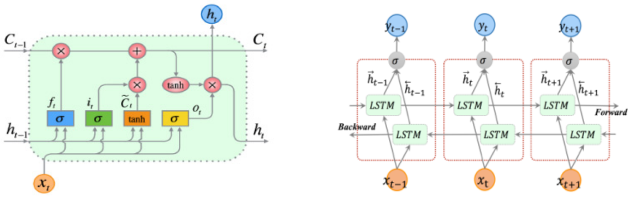
TABLE 1. **Student features during 1st and 2nd years.**

ID	KPI-1 features	ID	KPI-2 features
f1	Prep GPA $_{\text{Sem}=1,2}$	f7	GPA $_{\text{Sem}=3,4}$
f2	Prep Math 1 Gr	f8	Math 2 Gr
f3	Prep Phy 1 Gr	f9	Phy 2 Gr
f4	Prep Eng 1 Gr	f10	Eng 2 Gr
f5	Quizzes Gr	f11	Stat Gr
f6	Assignments Gr	f12	High school GPA
		f13	Overall GPA

* Prep: Preparatory first-year, GPA: grade point average, Gr: Grade, Phy: physics, Eng: English, Stat: statistics, Sem: Semester.

student retention rate. We will consider the student features in the prediction BLSTM problem as a sequence labelling model in which both the fetched student features and the output labels are constructed as sequences and saved into a feature vector. This is denoted by an input sequence KPI-1 = {f1, f2, f3, f4, f5, f6} and KPI-2 = {f7, f8, f9, f10, f11, f12, f13} with $n = 13$ feature size and the corresponding labels forming an output sequence $L = \{l_1, l_2, \ldots, l_t\}$, where t is time series. $F = \{f_{i=1}, f_{i=2}, \ldots, f_{i=n}\}$ stands for the input feature vector of semester S = {s1, s2, s3, s4}, whereas Yi stands for the corresponding output label.

C. TRAIN DATA USING BIDIRECTIONAL LONG SHORT-TERM MEMORY (BLSTM)



a) LSTM cell b) Bidirectional LSTM

FIGURE 2. **Main steps of a) LSTM cell and b) BLSTM architecture [24].**

BLSTM is regarded as one of the leading artificial recurrent neural networks (RNNs) widely used in classifications and predictions based on time series data. It can be used to process entire data sequences and retain them in compressed form using the sequence labelling method. BLSTM also performs well with long-term data dependencies [24], and for many sequence labelling tasks such as speech recognition and handwriting. This advantage inspired us to formulate the prediction problem from a sequence perspective. The previous step tries to characterize students with their features, and then, we need to train BLSTM method over the student features in both directions; namely, forward and backward with hidden states before concatenating the output from both directions. BLSTM demonstrated that it is efficient in various related works. The typical BLSTM structure [25] is depicted in **Fig. 2**. It is composed of basic LSTM units with each unit mapped between input sequence **X** and output sequence **Y**. This is given a student record is $\mathbf{X} = \{\mathbf{x1}, \mathbf{x2}, \ldots\ldots \mathbf{xn}\}$, its corresponding features is $\mathbf{F} = \{\mathbf{f1}, \mathbf{f2}, \ldots, \mathbf{f13}\}$, and label sequence is $\mathbf{L} = \{\mathbf{l1}, \mathbf{l2}, \ldots\ldots, \mathbf{ln}\}$. The mapping process goes through multiple gates including the input gate, forget gate, current memory cell, and output gate defined as it, $\mathbf{g_t}$, $\mathbf{c_t}$ **and** $\mathbf{o_t}$, respectively. Furthermore, each student $\mathbf{x_i}$ was mapped to a feature embedding vector $\mathbf{e_i^f} \in \mathbf{R^{d_f}}$, where $\mathbf{R^{d_f}}$ is the dimension of the student feature vector. Then, LSTM used hidden layer $\mathbf{h_i}$ which contains the mapped embedding vector $\mathbf{e_i^f}$, calculated as follows:

$$\mathbf{h_i} = \mathbf{M}\left(\mathbf{e_i^f}, \mathbf{h_{i-1}}\right), \tag{1}$$

Thus, $\mathbf{h_i} \in \mathbf{R^{d_h}}$, $\mathbf{d_h}$ is the size of hidden layers. Then, the mapping process continues at the $\mathbf{t^{th}}$ time step by the following equations:

$$\mathbf{i_t} = \sigma\left(\mathbf{W_{h_i}h_{t-1}} + \mathbf{W_{e_i}e_t^f} + \mathbf{b_i}\right), \tag{2}$$

$$\mathbf{g_t} = \sigma\left(\mathbf{W_{h_g}h_{t-1}} + \mathbf{W_{e_g}e_t^f} + \mathbf{b_g}\right), \tag{3}$$

$$\mathbf{\hat{c}_t} = \mathbf{tanh}\left(\mathbf{W_{h_c}h_{t-1}} + \mathbf{W_{e_c}e_t^f} + \mathbf{b_c}\right), \tag{4}$$

$$\mathbf{c_t} = \mathbf{g_t} * \mathbf{c_{t-1}} + \mathbf{i_t} * \mathbf{\hat{c}_t}, \tag{5}$$

$$\mathbf{o_t} = \sigma\left(\mathbf{W_{h_o}h_{t-1}} + \mathbf{W_{e_o}e_t^f} + \mathbf{b_o}\right), \tag{6}$$

$$\mathbf{h_t} = \mathbf{o_t} * \mathbf{tanh}(\mathbf{c_t}), \tag{7}$$

where $\mathbf{e_t^f}$ and $\mathbf{h_t}$ represent student feature and hidden state vectors at time **t**. $\sigma()$ and **tanh**() are the activation functions. $*$ denotes the element-wise product. **W** and **b** are BLSTM parameters representing the weight matrices and bias vectors. Finally, bidirectional LSTM employed the forward and backward student information in the following equations:

$$\overrightarrow{\mathbf{h_t}} = \mathbf{M}\left(\mathbf{e_t^f}, \overrightarrow{\mathbf{h_{t-1}}}\right), \tag{8}$$

$$\overleftarrow{\mathbf{h_t}} = \mathbf{M}\left(\mathbf{e_t^f}, \overleftarrow{\mathbf{h_{t+1}}}\right), \tag{9}$$

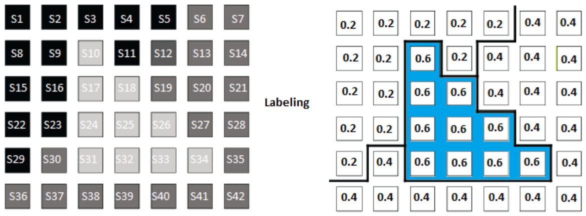
$\overrightarrow{\mathbf{h_t}}$ and $\overleftarrow{\mathbf{h_t}}$ are two hidden states in forward and backward directions saved into vector as an output of BLSTM as follows:

$$\mathbf{h_t} = [\overrightarrow{\mathbf{h_t}}, \overleftarrow{\mathbf{h_t}}], \tag{10}$$

The final out of this cell $\mathbf{y_t}$ is formed by $\mathbf{h_t}$, to represent the final sequence $\mathbf{L} = \{\mathbf{l_1}, \mathbf{l_2}, \ldots \mathbf{l_t} \ldots, \mathbf{l_n}\}$.

D. PREDICTION USING CONDITIONAL RANDOM FIELDS (CRF)

The main task of BLSTM is to extract and encode selected features of students based on two performance indicators:



a) Sample of students Si with their features.

b) Labeling students that have the same features with the same label in order to classify them.

FIGURE 3. Labeling student features using CRF.

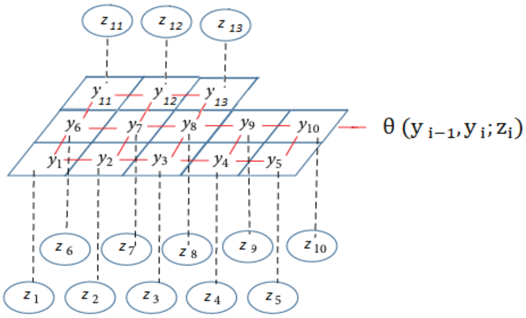


FIGURE 4. Conditional random field structure of students with 13 encoded features zi, where i = {1, . . . , n = 13}.

KPI − 1 and **KPI − 2**. It is hoped that it may provide clear indications of the differences between the encoded features of students who enrolled on the same course or in the same semester. This leads to the adoption of CRF [26] for sequence labeling to predict each student's label independently. It can then model the relationships between adjacent labels with a transition score and learn the interactions between pairs of features and labels with a state score as shown in Fig. 3. For instance, there are three types of students based on their features: 1) ongoing students; 2) students at risk of dropping out; and 3) unsurpassed university students. The ongoing or normal students have features weight labeled as 0.2 according to our experiments. Students at risk of dropping out feature weights are labeled as 0.4. The unsurpassed students have features weights labeled as 0.6 and highlighted in blue.

Our method was to define the input sequence for CRF as a hidden representation $\mathbf{Z_i} = \{\mathbf{z_i}$ **and** $\mathbf{z_i} = \mathbf{l_i}\}_{\mathbf{i=1}}^{\mathbf{N}}$, and the output sequence $\mathbf{Y_i} = \{\mathbf{y_i}\}_{\mathbf{i=1}}^{\mathbf{N}}$, then, CRF calculate the condition probability over all possible sequences $\mathbf{Y_i}$ given $\mathbf{Z_i}$ as follows:

$$\mathbf{P}\,(\mathbf{y} \mid \mathbf{z};\, \mathbf{W}, \mathbf{b}) = \frac{\prod_{\mathbf{i=1}}^{\mathbf{N}} \boldsymbol{\theta}_{\mathbf{i}}(\mathbf{y_{i-1}}, \mathbf{y_i}; \mathbf{z_i})}{\sum_{\mathbf{y' \in N}} \prod_{\mathbf{i=1}}^{\mathbf{N}} \boldsymbol{\theta}_{\mathbf{i}}(\mathbf{y'_{i-1}}, \mathbf{y'_i}; \mathbf{z_i})}, \tag{11}$$

where $\prod$ is the exponential operation and $\boldsymbol{\theta}_{\mathbf{i}}$ is the score function for the transition between the sequence pair $(\mathbf{y'}, \mathbf{y})$ for a given $\mathbf{z}$ as shown in **Fig. 4**. During training of the CRF model, we employed maximum likelihood estimation (MLE) as introduced in [27]. For a training pair $(\mathbf{z_i}, \mathbf{y_i})$, we maximize the loss function

$$\mathbf{Loss}(\boldsymbol{\theta}) = \sum_{\mathbf{i}}^{\mathbf{N}} \mathbf{log}(\mathbf{P}\,(\mathbf{y} \mid \mathbf{z};\, \mathbf{W}, \mathbf{b})) \tag{12}$$

During the decoding process, we search for the best sequence $\mathbf{y'}$ with the highest conditional probability as follows.

$$\mathbf{y'} = \mathbf{ArgMaxP}\,(\mathbf{y} \mid \mathbf{z};\, \mathbf{W}, \mathbf{b})) \tag{13}$$

At result, Conditional Random Fields is a discriminative model aimed to predict student retention where contextual information or state of the Neighbours affect the current prediction.

IV. EXPERIMENTAL RESULTS

The student data was collected from the University of Ha'il dataset. In our experiments, we trained student features using BLSTM to assign optimal weights. Then we applied the CRF model to mark student features with labels. Later we compared each student with their neighboring students to predict retention and graduation rates.

The university dataset consisted of 35,000 student records. We had two classes of student data based on the KPIs: first year and second year students. Two thousand first year students were selected from the preparatory dataset and 949 second years from the College of Computer Science and Engineering dataset. As a result, each student had 13 features.

We regarded the first and second semesters as a time series for predicting student retention rate and the third and fourth semesters as a time series for predicting student completion rate.

The principal objective was to investigate the number of students at risk of discontinuing their study as identified by dropout risk and completion within the time risk. In this paper, we selected registered students attending the College of Computer Science and Engineering (CCSE). Three College departments were included in our investigation: Computer Science (CS), Software Engineering (SE), and Computer Engineering (CE). We combined data from both students who were at risk and from those with unsurpassed records into a CCSE demographic dataset suitable for learning status prediction. Details of the CCSE dataset are presented in **Table 2**.

The preprocessed datasets were divided into two groups, 80% of which were used as training datasets and 20% of which were used as testing datasets. BLSTM was used as a deep learning model for time series predictions, the parameters and values of which are i displayed in **Table 3**.

Three evaluation metrics were used to examine the performance of our prediction model: precision P, recall R and $\mathrm{F_{score}}$. Their definitions are given below:

$$\mathbf{P} = \frac{\mathbf{TP}}{\mathbf{TP + FP}}, \tag{14}$$

$$\mathbf{R} = \frac{\mathbf{Tp}}{\mathbf{Tp + FN}}, \tag{15}$$

TABLE 2. Student statistics for CCSE 2020-2021.

Total number of students/department	CS	SE	CE	Total
	234	389	326	949
Total number of unsurpassed students/department Criteria: GPA<1	89	135	96	320
	Rate of unsurpassed students			
	38 %	35%	29 %	34%
Total number of unsurpassed students/department Criteria: completion time>4 years	21	55	64	140
	Rate of unsurpassed students			
	9%	14%	20%	15%

TABLE 3. Parameters set up for the BLSTM model.

Parameter	Value
Number of layers	3 − 4
Number of neurons in hidden units	128
Learning rate	0.001
Optimizer	Adam
Batch size	16
Epochs	300
Time step	2
Activation function	*Sigmoid*
Dropout rate	0.5
Computation mode	Parallel CPU

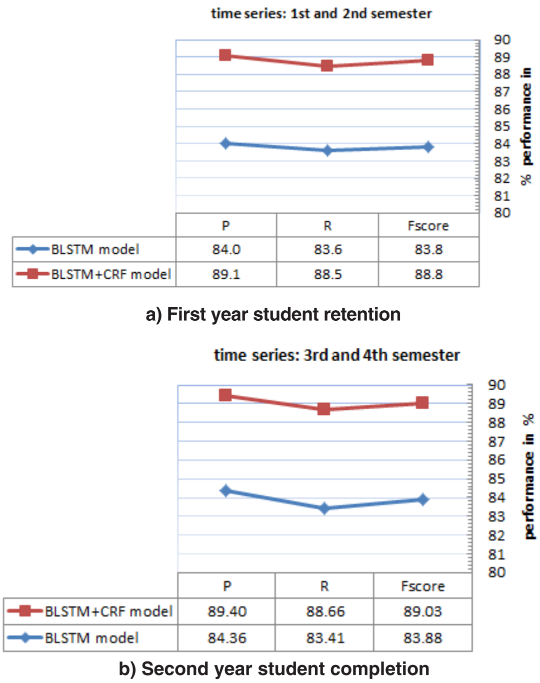
$$\mathbf{F_{score}} = \frac{\mathbf{2 * (R * P)}}{\mathbf{R + P}}, \tag{16}$$

where *TP* is the true positive, indicating that the retention status of the student is at risk, as is their predicted status. *FP* is a false positive, where the student's retention status is not at risk, but the predicted tag indicates that there is a risk. *FN* is a false negative, where the student's retention status is at risk but their predicted status suggests that there is no risk. We trained our prediction model over the first and second semesters, with the mean and standard deviation metrics shown in **Fig. 5**. The figure shows the performance of our model, the BLSTM with CRF, and the performance of another model using only the BLSTM, both of which were tested on the CCSE dataset.

Our solution achieved precision of 89.1, recall of 88.5 %, and an $\mathrm{F_{score}}$ of 88.8. The results of our method are promising in the second year and better than the performance of predicting student retention in the first year.

The most recent student data from the academic year 2020-2021 were used to evaluate the performance of our BLSTM + CRF method compared to that of the state-of-the-art prediction methods. **Table 4** compares the results obtained by Logistic Regression [28], Decision Tree [29], Random Forest [30], Naïve Bayes [31], Support Vector Machines [32], and Neural Network [33] methods based on the most intuitive measure/indicator of success: accuracy. It is simply defined as a ratio of correctly predicted student at risk retention status to the total number of students.

$$\mathbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + FP + FN + TN}}, \tag{17}$$



	P	R	Fscore
BLSTM model	84.0	83.6	83.8
BLSTM+CRF model	89.1	88.5	88.8

a) First year student retention

	P	R	Fscore
BLSTM+CRF model	89.40	88.66	89.03
BLSTM model	84.36	83.41	83.88

b) Second year student completion

FIGURE 5. Performance evaluation comparison of BLSTM + CRF model and BLSTM model only at time interval: a) first year student retention, and b) second year student completion.

TABLE 4. Comparison of performance evaluations.

Method	Accuracy	P	R	F_{score}
Logistic Regression [28]	0.93	0.79	0.98	0.90
Decision Tree [29]	0.90	0.98	0.71	0.85
Random Forest [30]	0.93	0.96	0.86	0.91
Naïve Bayes [31]	0.77	0.93	0.72	0.82
Support Vector Machines [32]	0.92	0.79	0.96	0.88
Neural Network [33]	0.88	0.86	0.89	0.88
BLSTM+CRF (our method)	**0.90**	**0.90**	**0.89**	**0.90**

V. CONCLUSION

The retention rates of students across their selected field of study and the graduation rates overall over a specified period are issues of concern to higher education organizations. Indeed, universities must be persistent in their efforts to improve student retention rates according to university teachers.

This paper has a dual focus. First, on the use of data preprocessing and deep learning algorithms to extract sensitive information about students. Their data in this case were extracted from a demographic data file. Second, on the utilization of key performance indicators with CRF methods as a classifier in the database. The unique contribution of this paper is: by employing these algorithms, we formulated a deep learning predictor model to investigate students' retention at four levels of study (over a two year period). Experimental results were obtained from a case study of the University of Ha'il. Regarding the BLSTM learning method, the results obtained with the predictive CRM approach indicated that prediction of student retention was possible with an accuracy of over 0.85 in most scenarios and with FP rates ranging from 0.05 to 0.10 in most cases.

Future work will utilize the ant colony optimization (ACO) method to enhance the structure of BLSTM cells, because it allows the researcher to determine the optimal weights of BLSTM connected cells. Furthermore, it can reduce the number of BLSTM cells required whilst at the same time improving predictive ability.

REFERENCES

[1] L. Thomas, ''Student retention in higher education: The role of institutional habitus,'' *J. Educ. Policy*, vol. 17, no. 4, pp. 423–442, 2002.

[2] E. R. Kahu, ''Framing student engagement in higher education,'' *Stud. Higher Educ.*, vol. 38, no. 5, pp. 758–773, 2013, doi: 10.1080/03075079.2011.598505.

[3] D. Raju and R. Schumacker, ''Exploring student characteristics of retention that lead to graduation in higher education using data mining models,'' *J. College Student Retention, Res., Theory Pract.*, vol. 16, no. 4, pp. 563–591, Feb. 2015, doi: 10.2190/CS.16.4.e.

[4] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, ''Predicting student dropout in higher education,'' in *Proc. ICML Workshop Data Good, Mach. Learn. Social Good Appl.*, New York, NY, USA, 2016, pp. 20–26.

[5] M. Tight, ''Student retention and engagement in higher education,'' *J. Further Higher Educ.*, vol. 44, no. 5, pp. 689–704, May 2020.

[6] O. Aljohani, ''A comprehensive review of the major studies and theoretical models of student retention in higher education,'' *Higher Educ. Stud.*, vol. 6, no. 2, pp. 1–18, 2016, doi: 10.5539/hes.v6n2p1.

[7] L. S. Hagedorn, ''How to define retention,'' in *College Student Retention: Formula for Student Success*. Westport, CT, USA: Greenwood Publishing Group, 2005, pp. 90–105.

[8] E. Kahu and J. Lodge, ''2018 special issue: Student engagement and retention in higher education,'' *Student Success*, vol. 9, no. 4, pp. 1–3, Dec. 2018, doi: 10.5204/ssj.v9i4.1141.

[9] T. Devasia, T. Vinushree, and V. Hegde, ''Prediction of students performance using educational data mining,'' in *Proc. Int. Conf. Data Mining Adv. Comput. (SAPIENCE)*, Mar. 2016, pp. 91–95, doi: 10.1109/SAPIENCE.2016.7684167.

[10] O. Aljohani, ''Analyzing the findings of the Saudi research on student attrition in higher education,'' *Int. Educ. Stud.*, vol. 9, no. 8, pp. 184–193, 2016, doi: 10.5539/ies.v9n8p184.

[11] C. Alexander, M. Kohnke, and A. Naginey, ''Undergraduate and graduate retention two concepts, one outcome,'' in *Proc. Nat. Symp. Student Retention*, Buffalo, NY, USA, 2009, p. 202.

[12] S. S. A. Tarmizi, S. Mutalib, N. H. A. Hamid, and S. A. Rahman, ''A review on student attrition in higher education using big data analytics and data mining techniques,'' *Int. J. Mod. Educ. Comput. Sci.*, vol. 11, no. 8, pp. 1–14, Aug. 2019, doi: 10.5815/ijmecs.2019.08.01.

[13] *Retention/Completion Summary Tables*, Amer. College Test., Iowa City, IA, USA, 2013.

[14] R. Styron, Jr., ''Student satisfaction and persistence: Factors vital to student retention,'' *Res. Higher Educ. J.*, vol. 6, p. 1, Mar. 2010.

[15] H. K. Swecker, M. Fifolt, and L. Searby, ''Academic advising and first-generation college students: A quantitative study on student retention,'' *NACADA J.*, vol. 33, no. 1, pp. 46–53, Jun. 2013, doi: 10.12930/NACADA-13-192.

[16] S. L. Britt, D. A. Ammerman, S. F. Barrett, and S. Jones, ''Student loans, financial stress, and college student retention,'' *J. Student Financial Aid*, vol. 47, no. 1, p. 3, 2017. [Online]. Available: https://ir.library.louisville.edu/jsfa/vol47/iss1/3

[17] D. Delen, ''Predicting student attrition with data mining methods,'' *J. College Student Retention, Res., Theory Pract.*, vol. 13, no. 1, pp. 17–35, May 2011, doi: 10.2190/CS.13.1.b.

[18] I. E. Allen and J. Seaman, *Changing Course: Ten Years of Tracking Online Education in the United States*. Newburyport, MA, USA: Sloan Consortium, 2013.

[19] W. E. Boston and P. Ice, ''Assessing retention in online learning: An administrative perspective,'' *Online J. Distance Learn. Admin.*, vol. 14, no. 2, pp. 133–137, 2011. [Online]. Available: http://www.westga.edu/distance/ojdla/summer142/boston_ice142.html

[20] J. Gaytan, ''Comparing faculty and student perceptions regarding factors that affect student retention in online education,'' *Amer. J. Distance Educ.*, vol. 29, no. 1, pp. 56–66, Jan. 2015, doi: 10.1080/08923647.2015.994365.

[21] M. H. Windham, M. C. Rehfuss, C. R. Williams, J. V. Pugh, and L. Tincher-Ladner, ''Retention of first-year community college students,'' *Community College J. Res. Pract.*, vol. 38, no. 5, pp. 466–477, May 2014.

[22] E. Osmanbegovic and M. Suljic, ''Data mining approach for predicting student performance,'' *Econ. Rev., J. Econ.*, vol. 10, no. 1, pp. 3–12, 2012. [Online]. Available: http://hdl.handle.net/10419/193806.

[23] E. Rahm and H. H. Do, ''Data cleaning: Problems and current approaches,'' *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, Dec. 2000.

[24] Z. Cui, R. Ke, Z. Pu, and Y. Wang, ''Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values,'' *Transp. Res. C, Emerg. Technol.*, vol. 118, Sep. 2020, Art. no. 102674, doi: 10.1016/j.trc.2020.102674.

[25] T. Thireou and M. Reczko, ''Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 3, pp. 441–446, Jul./Sep. 2007, doi: 10.1109/tcbb.2007.1015.

[26] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, ''Conditional random fields as recurrent neural networks,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1529–1537.

[27] I. J. Myung, ''Tutorial on maximum likelihood estimation,'' *J. Math. Psychol.*, vol. 47, no. 1, pp. 90–100, 2003, doi: 10.1016/S0022-2496(02)00028-7.

[28] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, and C. D. Kloos, ''Prediction in MOOCs: A review and future research directions,'' *IEEE Trans. Learn. Technol.*, vol. 12, no. 3, pp. 384–401, Jul. 2019, doi: 10.1109/TLT.2018.2856808.

[29] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, ''Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization,'' *Comput. Hum. Behav.*, vol. 58, pp. 119–129, May 2016.

[30] W. Wang, H. Yu, and C. Miao, ''Deep model for dropout prediction in MOOCs,'' in *Proc. 2nd Int. Conf. Crowd Sci. Eng.*, 2017, pp. 26–32, doi: 10.1145/3126973.3126990.

[31] E. M. Queiroga, J. L. Lopes, K. Kappel, M. Aguiar, R. M. Araújo, R. Munoz, R. Villarroel, and C. Cechinel, ''A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course,'' *Appl. Sci.*, vol. 10, no. 11, p. 3998, Jun. 2020, doi: 10.3390/app10113998.

[32] T. R. Hagedoorn and G. Spanakis, ''Massive open online courses temporal profiling for dropout prediction,'' presented at the IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI), Boston, MA, USA, Nov. 2017.

[33] M. Youssef, S. Mohammed, E. K. Hamada, and B. F. Wafaa, ''A predictive approach based on efficient feature selection and learning algorithms' competition: Case of learners' dropout in MOOCs,'' *Educ. Inf. Technol.*, vol. 24, no. 6, pp. 3591–3618, Nov. 2019.

DIAA ULIYAN received the Ph.D. degree in computer science from the University of Malaya, Kuala Lumpur, Malaysia, in 2016. Since 2019, he has been with the College of Computer Science and Engineering, University of Ha'il, Hail, Saudi Arabia, as an Assistant Professor. His research interests include deep learning, computer vision, and information security.

ABDULAZIZ SALAMAH ALJALOUD received the M.Sc. and Ph.D. degrees from the University of New England, Australia, in 2015 and 2018, respectively. He is currently working as an Assistant Professor with the College of Computer Science and Engineering, University of Ha'il, Saudi Arabia. He is also working as the Vice Dean of the Quality Assurance Unit, University of Ha'il. His primary research interests include machine learning and signal processing.

ADEL ALKHALIL received the Ph.D. degree from Bournemouth University, Poole, U.K. He joined the College of Computer Science and Engineering, University of Ha'il, Ha'il, Saudi Arabia, as an Assistant Professor. His research interests include software evolution for mobile and cloud computing systems, decision support systems, and knowledge-based systems.

HANAN SALEM AL AMER received the Ph.D. degree in mathematics curriculum and teaching methods from King Abdulaziz University, Saudi Arabia. She joined the Faculty of Education, Department of Curriculum, University of Ha'il, as an Associate Professor of mathematics curriculum and teaching methods, where she is currently the Deputy Minister for private general.

MAGDY ABD ELRHMAN ABDALLAH MOHAMED is currently an Associate Professor of educational planning and a Quality and Accreditation Expert. He is also the Director of the Department of Planning, Studies, and Institutional Excellence, Deanship of Quality and Development, University of Ha'il.

AZIZAH FHAD MOHAMMED ALOGALI is currently a Consultant at Saudi Standards, Metrology and Quality Organization, Saudi Arabia. She is also the Principal Consultant at Nagarro Information Technology Company, and the Administration Director of the American Society for Quality, Saudi Arabia.