Predicting At-Risk Students' Performance Based on LMS Activity using Deep Learning

Amnah Al-Sulami[1], Miada Al-Masre[2], Norah Al-Malki[3]
Information Technology Department, King Abdulaziz University, Jeddah, Saudi Arabia[1,2]
Modern Languages and Literatures, King Abdulaziz University, Jeddah, Saudi Arabia[3]

Abstract—**It is of great importance for Higher Education (HE) institutions to continuously work on detecting at-risk students based on their performance during their academic journey with the purpose of supporting their success and academic advancement. This is where Learning Analytics (LA) representing learners' behaviour inside the Learning Management Systems (LMS), Educational Data Mining (EDM), and Deep Learning (DL) techniques come into play as an academic sustainable pipeline, which can be used to extract meaningful predictions of the learners' future performance based on their online activity. Thus, the aim of this study is to implement a supervised learning approach which utilizes three artifcial neural networks (vRNN, LSTM, and GRU) to develop models that can classify students' final grade as Pass or Fail based on a number of LMS activity indicators; more precisely, detect failed students who are actually the ones susceptible to risk. The three models alongside a baseline MLP classifier have been trained on two datasets (ELIA 101-1, and ELIA 101-2) illustrating the LMS activity and final assessment grade of 3529 students who enrolled in an English Foundation-Year course (ELIA 101) taught at King Abdulaziz University (KAU) during the first and second semesters of 2021. Results indicate that though all of the three DL models performed better than the MLP baseline, the GRU model achieved the highest classification accuracy on both datasets: 93.65% and 98.90%, respectively. As regards predicting at-risk students, all of the three DL models achieved an = 81% Recall values, with notable variation of performance depending on the dataset, the highest being the GRU on the ELIA 101-2.**

Keywords—*Predict at-risk student; artificial neural network; learning management system; and educational data mining*

I. Introduction

Educational Data Mining (EDM) is currently an exciting field of Data Mining (DM) which deals with investigating Educational Big Data and Learning Analytics (LA) with the purpose of conceptualizing models that can be effectively used in enriching the learners' experiences and augmenting educational institutions' academic offerings and efficiency. Traditionally, EDM applies DM, Machine Learning (ML), and statistical methods to identify patterns in large educational data [1]. Many researchers established the effectiveness of using DM techniques in the educational field, especially, in domains that are crucial to learners' progress, engagement and performance [2] [3].

The previously cited metrics of the students' learning journey, and others, are closely connected to how decision-makers and educators are preoccupied with proactively identifying at-risk students based on their behaviours and performance in educational environments. Currently, EDM, in combination with ML and Deep Learning (DL) techniques have greatly impacted academic decision-making in terms of robustness, accuracy, and sustainability because mining LA to discover information about students' learning have led to many preemptive measures that support learners' success and advancement [4] [5].

In E-learning environments, LA is oftentimes representative of a user's behaviour inside an institutional Learning Management System (LMS). Theoretically, and in educational contexts, users' behaviour denotes the interactions performed by users inside a website, a mobile app, or a system which can be monitored through analytics tools. The detection of a user's behaviour is often dependent on features which demonstrate the amount, continuation, and emphasis of user activities [6]. These behaviours are significant factors in the evaluation of why a certain user interacts with the system in a specific way, how to proactively predict these behaviours; consequently, detect their impact on the endpoint of the process. In an educational setting, students' behaviour data represent the activities and learning interactions; ideally, within an E-learning platform such an LMS, where there are tools that can help in collecting and storing such data for further analysis. The features of this data can be, for example, course accesses, submissions, clickstream, time-series data, videos, lectures, assessments, discussion forums, and even live video discussions through the internet [7]. The features of this data are usually analysed to predict students' academic achievement, develop recommendation systems, analyse students' behaviour, re-design courses, and identify at-risk students.

Using LA with a combination of EDM, ML and DL to detect at-risk students in Higher Education (HE) Institutions has become the focus of contemporary research, where identifying at-risk learners is often, if not always, associated with demographic, social, psychological, or cultural factors both inside and outside the institution and their impact on final grades in courses and GPAs, or outcomes in programs [8] [9]. Attention to LA is, as well, crucial to a rounded understanding of learners who are susceptible to risk. Since the early 2000s, we discern a change in demarcating the scope of at-risk students, which is motivated by the mainstream use of online environments where LA has become another indicator of learners' behavioural activity in educational settings, as well as the possibility of utilizing ML, generally EDM, techniques to determine risk factors and outcomes [10]. We observe, however, the scarcity of research that addresses practical methods for detecting at-risk students based on DL algorithms in HE contexts.

The current study is primarily motivated by the need for prior discovery of at-risk students through examining their user behaviour in an online learning offering during COVID-

19 in King Abdullaziz University (KAU), which is originally delivered in the same format even before the transformation to Distance Education throughout the pandemic lockdown. The ELIA 101 course is designed and taught in a blended format to Foundation-Year students in KAU including a number of assessments that are submitted via the official LMS, Blackboard. The ultimate aim of at-risk learners' identification is to improve their performance by giving them the opportunity to enhance their achievement and avoid dropout or being academically dismissed from the programs. We assume that, based on the automated predictive modelling of Foundation-Year students' online interactions data, KAU can progressively improve the engagement of low-performing learners, predict students' final grade indicative of their Pass/Fail performance, and prevent their dropout from the course.

Therefore, the aim of this study is to 1) create two datasets which represent the main features of assessment design in ELIA 101 alongside other meaningful online activity attributes, 2) develop three Artificial Neural Networks models that classify students based on their final grade; consequently, detect at-risk students enrolled in the ELIA 101 English course, and 3) evaluate the performance of the models focusing on their accuracy and effectiveness. Generally, we will be investigating answers to the following research questions:

1) Which DL network achieves the highest accuracy in detecting students' at-risk status (Fail)?
2) Which DL network achieves the highest accuracy in classifying students Pass/ Fail status in the course?

Ideally, this study's contributions can be outlined as follows:

- The collection and pre-processing of two datasets for training the at-risk students prediction models.
- The development of three neural networks, i.e., vanilla Recurrent Neural Network (vRNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), to predict at-risk students in ELIA 101 Foundation-Year English course at KAU.

The rest of this paper is organized as follows: Section II is a literature review of ML and DL methods used in students' performance, and at-risk status prediction. Section III, describes the research methodology. Section IV, presents the findings, discussion and limitations of our research. Section VI, is the conclusion with reference to future work.

II. Review of Literature

Predicting students' final grade (including the identification of their at-risk status) is representative of their course performance, and help academic institutions support their students' success, as well as encourage learners to change their study patterns and get better grades.

From an ML perspective, there has been numerous research which experimented with various techniques to predict learners' final grade, and detect their at-risk status using supervised learning methods with a specific implementation of binary classification. Among the studies which considers ML methods is Macarini in which thirteen datasets have been created from 89 students' activity on the Moodle LMS. The classification algorithms used to classify these datasets are K-Nearest Neighbors (K-NN), Multilayer Perceptron (MLP), Random Forest (RF), AdaBoost, and Naive Bayes. Results show that a combination of the AdaBoost with dataset2 and dataset5 have performed better than the rest of the models [11].

Similarly, Kumari, assessed the behavioural features which could be effective in enhancing students' performance. The data is collected from the LMS for 500 students using the experience API (xAPI). Their model utilized ML algorithms like Iterative Dichotomiser 3 (ID3), K-NN, naive Bayes, Support Vector Machine (SVM). The algorithms have been implemented on the Waikato Environment for Knowledge Analysis (WEKA), where the ID3 achieved a higher accuracy than the other methods (=90%) [12].

Besides, Karthikeyan have assessed students' performance by proposing a Hybrid Educational Data Mining (HEDM) model. This model combines the effectiveness of naive Bayes and the J48 Classifier classification technique. The model has been tested against an online dataset and achieved an 98% accuracy [13].

The investigation of at-risk students' activity metrics and impact on their final grade have been also studied with a combination of ML and basic DL techniques like ANNs. ML and DL algorithms have been developed in Howard, on a dataset consisting of 136 students' LMS activity, the researchers implemented RF; XGBoost; Bayesian Additive Regression Trees (BART); Principal Components Regression (PCR); SVM; Multivariate Adaptive Regression Splines (MARS); neural network; and K-NN. They used the actual final grade as the main variable to which they compared the predicted one. The Mean Absolute Error (MAE) was calculated; and they reported that PCR had the lowest MAE value 6.5. The researchers found that the best time to expect at-risk students was during weeks 5/6 [14].

In Hung the data representing 12,869 students has been collected from a K-12 virtual school in the northern USA. The algorithms used for the model are SVM with the sigmoid kernel, SVM with polynomial kernel, SVM with gaussian radial basis function, RF, and ANN. The DL model achieves better performance than the ML ones by correctly identifying 51% of at-risk students with 86% accuracy [15].

Besides, Altabrawee, utilizes both ML and DL models to predict students' performance in a computer science subject at Al-Muthanna University, which is tested on data representing the user behaviour of 161 students. The researchers design an ANN, Naïve Bayes, decision tree, and logistic regression models. As indicated by results, the ANN model achieved a 77% accuracy, higher than the other models [16].

The early at-risk detection of students' performance enables them to improve their learning strategies. For example, Sultana, develop models to warn learners who have low performance issues based on their cognitive and non-cognitive competences to decrease their dropout. The non-cognitive features include: Time-management, Self-concept, Realistic-Self-appraisal, and Community support. The dataset used in this research represent the user activities of 778 students collected from different universities and online repositories. The researchers have applied logistic regression, decision tree, naive Bayes, and an ANN. Results indicate that certain combinations of cognitive and non-cognitive features improved the model performance. For example, a combination of cognitive features, Leadership and Realistic-Self-appraisal data trained with a naive Bayes model has resulted in the highest accuracy value, 65%. Similar cognitive, non-cognitive combinations have achieved the same accuracy with the naive Bayes model as well [17].

With a specific focus on DL methods, Aydoğdu uses an ANN for student final performance prediction. The dataset, used in this research, comes from the activity stream of 3518 students. The model achieves an accuracy of 80% [18]. In the same manner, the researchers in Hussain, design a method to predict students' results, which is tested on a dataset representing the user behaviour of 10140 students. The results demonstrate the effectiveness of the RNN which achieves an accuracy of 95% [19]. Utilizing an RNN as well, He, proposed a novel joint RNN-GRU neural networks that predicts at-risk students using OULAD. Three algorithms are considered as baseline models: vRNN, GRU, and LSTM. The findings show that simple techniques such as GRU and vRNN have better outcomes than the relatively complex model of LSTM. The joint model successfully predicted at-risk students at the end of the semester and obtained over 80% accuracy [7].

Prior studies examining student behavior mainly on an online or MOOC dataset, this study uses real students' datasets from the Blackboard LMS, the adopted LMS in most Saudi higher education institutions.

Moreover, the extracted LA data in the rest of the studies, which use a real student's dataset, is for a limited number of students except in [19], which has 10,140 learners. This study worked with two datasets created from data values representing 3,529 learners.

Therefore, this study extends previous research by providing an effective solution for predicting students' pass/ fail status and identifying at-risk students (fail) by implementing DL models based on individual student behavior in LMS.

III. Methodology

The methodology adopted for this research consists of four key phases: data collection, data pre-processing, development of prediction models, and evaluation (see Fig. 1).

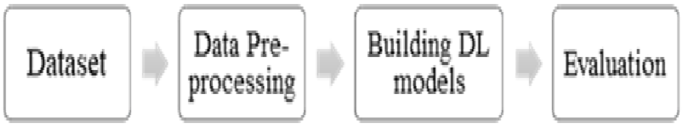

Fig. 1. Methodology pipeline.

More specifically, the process of the proposed models' pipeline could be illustrated in Fig. 2.

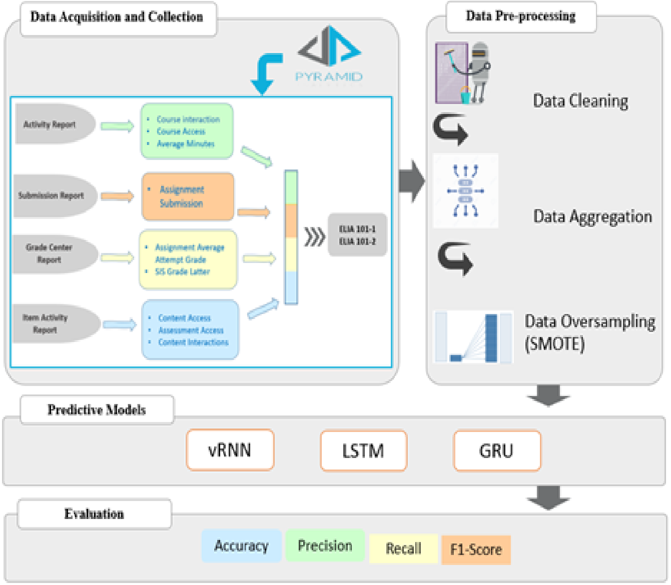

Fig. 2. The proposed models' pipeline.

A. Data Collection

The datasets used in this study are retrieved via a collaboration with the Deanship of E-Learning and Distance Education at KAU using the official Learning Analytics system, A4L which supports the Blackboard LMS. This service facilitates the extraction of big educational data; specifically, in our case, detailed reports on students' course activity (overall interactions), submissions, grade centre assessments, and activity on item level alongside their final grade in the course.

The A4L system has two interfaces, one for the institution and the other is user-centric, accessible via the LMS.

1) The Institution Interface: A4L records and displays students' LMS behaviour and interactions sliced by multiple data dimensions. Data measures in the form of frequencies, averages, and percentages can be extracted. Examples of dimensions include: Course, Advisor, SIS Major, SIS Student College, SIS Student Level, Student Risk Profile, Term, Time Series, etc. Examples of measures are Items Accessed, Assessments Accessed, Course Accessed, Content Accessed, etc.

2) The User Interface: Both instructors and students have a view-only permission to the A4L reports, which allow them to compare their performance to other users and follow their progress.

From the A4L solution described above, two datasets have been extracted. Both datasets are comprised of a total of 3529 students' activity data from an English Foundation-Year course (ELIA 101), which is delivered during the first and second semesters of 2021. The first dataset (ELIA 101-1) is from the Spring term run of (ELIA 101) and consists of 75,971 records representing (2386) students. (2322) of those students passed the course, whereas (64) of them failed. The second dataset (ELIA 101-2) is the Fall run of the (ELIA 101) course and included 26,291 records of 1,143 students. 1,137 of those students passed the course, and 6 of them failed.

TABLE I. DATASETS STRUCTURE

User Id	Asseement Access	Content Access	Course Access	Course Interactions	Course Item Interactions	Avg. Minutes	Assignment Submission Count	Assignment Attempt Grade
STD_1	0	3	2	4	3	117.717	1	100
STD_1	6	1	2	87	7	63.3	1	100
STD_1	3	12	3	96	16	74.367	1	100
STD_1	0	1	1	10	1	209.05	1	100
STD_1	0	0	1	1	0	0.217	1	100

TABLE II. FEATURES DESCRIPTION

Reports	Features	Description
Activity	Course_Access	A count of students' access per the course.
	Course_Interactions	A count of students' interactions per the course.
	Avg_Minutes	Average minutes that the students spent per course
Subnission	Assignment_Submission	A count of students' submissions per a specific assignment.
Grade Center	Assignment_Avg_ Attempt_Grade	Average students' grade per a specific assignment.
	SIS_Grade_Letter	Corresponding final course grade for the students.
Item Activity	Assessment_Access	A count of students' access per a specific assessment.
	Content_Access	A count of students' access per a specific content.
	Course_Item_Interactions	A count of students' interactions per a specific item.

Ideally, this course is delivered, at least, twice a year. The dataset is basically extracted as a report made up of several online activity indicators. Both datasets included 9 students' activity metrics as shown in Table I. The description of these metrics is listed in Table II.

The following inclusion and exclusion criterion for dataset creation have been observed and implemented:

1) Inclusion criterion:
   - The extracted reports included features (measures) representative of 1) the basic elements of instructional design, specifically, assessments and engagement indicators like interaction, as well as, 2) potential risk factors, which might have an impact on the instructional context of ELIA 101 as delivered in the English Language Institute in KAU, where, for example, assignments are delivered weekly via Blackboard as per required by the official Course Specification for this course [20]. Other assessment types like a Final Speaking Exam and a Final Writing Exam are used but are summative assessment tools like the Final Exam, and what we are interested in are formative assessments conducted during the semester. According to the instructional strategy of the course, forums and other collaborative tools are not used as assessment methods, henceforward, data related to them are not included [20].
   - Comparability is ensured through extracting data about a definable student cohort (Foundation-year students enrolled in the English Course, ELIA 101.
   - The data was extracted from both course and students' perspectives for 3529 students.
   - The researchers verified that the course included actual activities so that the extracted reports reflect actual user behaviour.
   - Extracted reports coverage is of the first and second semesters of 2021, where KAU migrated to the online learning platform to ensure the continuity of its academic offering during the COVID-19 lockdown.

2) Exclusion criterion:
   - Measures which relate to logins were excluded, because they display data representative of all the courses a student is enrolled in, not just ELIA 101.
   - Measures which perform complex statistical operations on the data (change rate, moving averages) were as well not considered.

B. Data Pre-processing

1) Data Cleaning: Data cleaning is a data pre-processing technique that is used to improve the quality of the data. This process ensures that there is no data nosiness or inconsistency, thus eliminating what researchers consider "garbage" [21]. In this step, the researcher cleans up the data by removing students with undefined grade schemas in the "SIS_Grade_Letter" column. For example, values like "NF", "No SIS Match", "W", "XX", and "No Recorded Grade" were removed as they have no relevant reference in the KAU grades schema except for (W) which indicates a student who dropped from the course or the program. Moreover, the grade above the actual (100), reported usually as a percentage mean of assessment grade, is removed from the "Assigment_Avg_Attempt_Grade" and "SIS_Grade_Letter" columns, which sometimes reflect an addition of an extra grade by the instructor that disrupts the percentage representation. However, only a few cases of these instances have been found (only 169 grade representations in the two datasets). Similarly, the "SIS_Grade_Letter" "DN" and "F" grade marks have been encoded as zeros because these symbols represent students who failed the course either due to their non-attendance or getting a grade lower than 60 out of 100.

2) Data Aggregation: The original dataset consisted of 75,971 entries for ELIA 101-1, and 26291 for ELIA 101-2), ranging from two to sixteen rows per student ID indictive of their level of interaction with the course. This necessitates that we find a method by which data representation becomes uniform for all students. So, conditional aggregation of data points based on an index (Student ID) has been performed through computing the mean of all the interaction values per student. The 'mean' value was used for aggregation because, unlike the 'count' and 'median' values, it does not affect the student's final grade, which we have noticed during experimentation with various data aggregation methods.

3) Data Imbalance: The number of students who have failed both ELIA 101-1 (64) and ELIA 101-2 (6) is small compared to the number of passing students in both datasets: 2322, 1137, respectively. This indicates (as Fig. 3 and 4 show) that there is a data imbalance problem that needs to be addressed so that the minority class, in this case "Fail", is not misrepresented, or affect the performance of the model. Therefore, data balancing strategies are applied to avoid a lower performance by DL methods which usually expect a balanced class distribution [22].

One method to overcomes data imbalance is Oversampling which increases the instances of the class with fewer numbers. As a first step, Random Oversampling (ROS) has been applied to the ELIA 101-1 and ELIA 101-2 datasets, but it only duplicates the data of the minority class which causes the models to overfit [23].

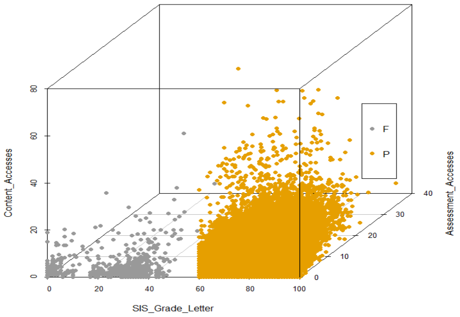

Fig. 3. ELIA 101-1 Pass/Fail imbalance.

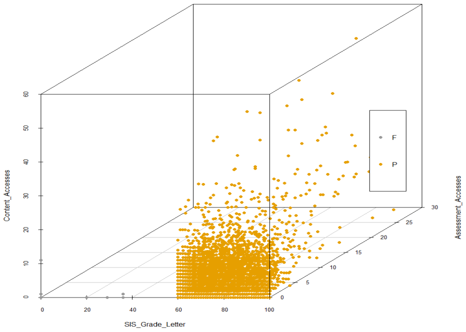

Fig. 4. ELIA 101-2 Pass/Fail imbalance.

To overcome overfitting, the Synthetic Minority Oversampling Technique (SMOTE) is employed [24]. where it populates the minority class instance (Fail) in the datasets by generating a synthetic sample that selects one of the k nearest neighbours of the feature in the feature space, and in the process, draws a line between the examples in the feature space, while defining a point along that line to generate the sample.

Oversampling the two datasets with SMOTE resulted in balanced datasets. ELIA 101-1, upon oversampling consisted of 4772 samples with equal Pass/Fail count, and ELIA 101-2 included 2286 samples with an equal distribution of the Pass/Fail grading schema.

4) Data Labelling: A Target column is added to the oversampled dataset, which translates the “SIS_Grade_Letter” grade value into 1 or 0 binary, indicating a Pass/Fail status, respectively. Students who score less than 60 on their final grade are classified as failing students, whereas students scoring 60 or above are recorded as pass.

5) Training and Testing Split: The dataset was split into training and testing sets with an 80/20% ratio. The data splits have been stratified by the target column so that neither the training nor testing sets be made completely of either one of the students’ status indicators 0 or 1. Consequently, the proportion of instances of each class (Pass/Fail) in each subset (Train/Test) is almost equal to that in the original dataset. We assume, as research indicates, that stratification improves the model’s performance as well as contributes to avoiding both bias issues related to variance.

TABLE III. MODELS’ HYPERPARAMETERS

Models	Layers	Optimizer/ Loss/ Metric	Epochs	Batch Size
vRNN	- Input (128, relu) - Hidden (64, relu) - 2 Hidden (32, relu) - Output (1, sigmoid)	- Adam (learning rate=0.001) - binary_crossentropy - accuracy	100	32
LSTM	- Input (128, relu) - Hidden (64, relu) - 3 Hidden (32, relu) - Output (1, sigmoid)			
GRU	- Input (256, relu) - Hidden (128, relu) - 5 Hidden (64, relu) - Output (1, sigmoid)			

C. Building the DL Models

The main objective of this research has been to adopt a supervised learning approach to develop DL models that are capable of predicting the presence of at-risk students depending on their LMS interaction with the various course components that made up the final grade (Pass/Fail). More specifically, we performed a binary classification of the two datasets representing the ELIA 101 Course based on students’ final grade. As we are considering a dataset with multiple predicators, we opted for developing and comparing the performance of classification models that are capable of prediction based on multiple indicators. To achieve this objective, we experimented with three deep learning models (vRNN, LSTM, and GRU).

All deep learning models have been trained and tested using Python 3 and TensorFlow 2.6.0. The three models’ hyperparameters are set to the ones illustrated in Table III. In order to avoid the overfitting problem, early stopping is used for all models [25].

1) Baseline Model: An initial implementation of a baseline neural network has been attempted, where the results are later used as a reference point of comparison to the proposed models. The objective is to generally investigate the efficacy of our deep learning approach while using baseline structures that can be improved on.

a) Multilayer Perceptron (MLP):
An MLP is one of the simplest representations of neural networks. It is a class of Feedforward ANNs which is composed of multiple layers of neurons that are connected through directed connections to the neurons of each subsequent layer [1]. There are three layers of neurons, including the input, hidden, and output layers. In its hidden and output layers, MLP uses sigmoid functions to predict probabilities. As part of the training process, MLP adjusts the weights iteratively by learning through a backpropagation function to produce good results [9]. An MLP with a hidden layer demonstrates a nonconvex loss function which results in multiple local minima [26]. Moreover, the decision process in an MLP is made in relation to the immediate input. It does not have memory of past or future input.

2) Proposed Models:

a) Vanilla Recurrent Neural Networks (vRNN):
RNNs are a type of artificial neural network which consists of connected nodes in a directed or undirected graph [27]. In RNNs, the information loops through the middle-hidden layer. The input is passed to the input layer, processes it, then passes it to the middle layer. The middle layer usually consists of multiple hidden layers, each with its own activation functions and weights. Unlike MLP, which is a Feedforward network considering only the current input, RNNs implements a feedback loop that ensures information cycles, which means that unlike feedforward connections, RNNs can also have connections that feed information to the previous or same layer [1]. Vanilla RNN is the standard RNN, it passes input as well as a hidden state through a single tanh layer [7]. In this research, the vRNN model is constructed of one input layer with 128 units, one hidden layer with 64 units, 2 hidden layers with 32 units, and the output fully connected layer. All layers use a Rectified Linear Unit (ReLU) activation function, except for the output layer which uses a 'sigmoid' activation. The model training stopped at 37 epochs on the (ELIA 101-1) dataset, and on 35 on the (ELIA 101-2) dataset.

b) Long Short-Term Memory (LSTM):
The main issue with vRNNs is the problem of vanishing gradients, which suffers an exponential decrease as the neural network back-propagates, which slows up the learning process in the final layers of the RNN [1]. Therefore, an LSTM model is developed to counteract the limitation of the vRNNs special architecture. The memory cell concept is introduced in LSTM architecture, which enables long-term dependency learning. As a function of its inputs, the memory cell temporarily holds its value, and is composed of three gates that regulate how information flows. Basically, there is an input gate which regulate when new information accesses the memory cell; a forget gate which manages the time limit for storing information in the cell, thus permitting new information to flow in; and the output gate that manipulates when the stored information is to be utilized by the processor [1]. Ultimately, this speeds the learning process as well as retains its information. Our LSTM model is constructed of an input layer with 128 neurons, a hidden layer contains 64 neurons, two hidden layers that include 32 neurons. Finally, the fully connected layer with one neuron was applied. Whereas all layers make use of a ReLU activation function, the output layer applies a 'sigmoid' activation. The model training stopped at 79 epochs with the (ELIA 101-1) dataset, and on 51 with the (ELIA 101-2) dataset.

c) Gated Recurrent Unit (GRU):
LSTM relies on using more training parameters, therefore uses more memory and executes slower than other models. So, for addressing this, we developed a GRU model, which is a simplified version of an LSTM [1]. Its hyperparameters are fewer in comparison to the LSTM as it consists of two gates in place of three, a reset and update gates. The GRU's update gate is a merge up of the LSTM's input and forget gates [7].

In this research, GRU is constructed with 7 layers; the input layer has 256 neurons, one hidden layer made of 128 neurons, five hidden layers including 64 neurons, and the output layer which has 1 neuron. All layers use an activation function, a ReLU, except the output layer which uses a 'sigmoid'. The model training stopped at 88 epochs when trained on the (ELIA 101-1) dataset and on 96 when trained on the (ELIA 101-2).

D. Evaluation

To evaluate the performance of the baseline and proposed models, we used the following metrics:

- Accuracy: is a metric that helps determine the percentage of correctly categorized instances in relation to the total of those instances based on the following Eq. 1 [9]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Recall: is a metric that is used to evaluate model overall. It computes the percentage of the correctly classified true positives using the following equation 2 [9]:

$$Recall(TPR) = \frac{TP}{TP + FN} \tag{2}$$

- Precision: is a metric that is used to assess the accuracy of the model. It reflects the percentage of true positives to those instances listed as positive by the classifier employing the following Eq. 3 [9]:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- F1-score: is a metric that computes the average of precision and recall, and it is useful in cases where the performance of different classifiers is to be compared. The following Eq. 4 represents the F1-core computation process [9]:

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

IV. Results and Discussion

The aim of this research is set initially to predict at-risk students based on their behaviour inside the ELIA 101 course in two datasets (ELIA 101-1) and (ELIA 101-2), which has been delivered via KAU's official LMS, Blackboard. Therefore, the proposed models' performance (vRNN, LSTM, GRU) are evaluated and compared to each other as well as the baseline model (MLP). Table IV shows the results of the various models' predictions on the datasets.

The MLP, as baseline, achieved an accuracy of 84% on the (ELIA 101-1) and 91.65% when tested on the (ELIA 101-2) dataset, which sets it as the model with the lowest performance in classifying students' Pass/Fail status. The GRU model, on the other hand, achieved an accuracy of 94.40% on the (ELIA 101-1) dataset and 98.02% on the (ELIA 101-2) which is considered the highest set of values among baseline and proposed models. The GRU, as well, achieved the best f1-scores (= 94.25% and 97.99%, respectively).

According to the results outlined above, a comparison of all predicted models is provided in Fig. 5 and 6.

Moreover, Fig. 7, 8, 9, 10, 11, and 12 illustrate the loss and accuracy of training and validation data per each model

TABLE IV. Performance Results for the Models on the Two Datasets

Models	Dataset	Accuracy	Precession	Recall	F1_Score
MLP	ELIA 101 -1	83.85	82.98	85.13	84.04
	ELIA 101 -2	91.65	98.96	84.14	90.95
vRNN	ELIA 101 -1	84.50	86.53	81.68	84.04
	ELIA 101 -2	96.92	**100**	93.83	96.82
LSTM	ELIA 101 -1	93.22	97.62	88.58	92.88
	ELIA 101 -2	89.24	92.72	85.13	88.76
GRU	ELIA 101 -1	**93.65**	**97.87**	**89.22**	**93.35**
	ELIA 101 -2	**98.90**	**100**	**97.79**	**98.89**

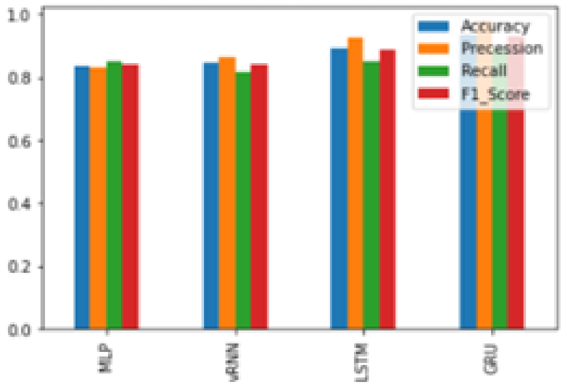

Fig. 5. Comparison of performance results for the models on ELIA 101-1.

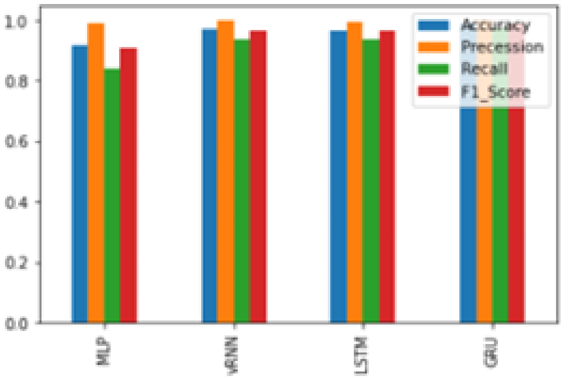

Fig. 6. Comparison of performance results for the models on ELIA 101-2.

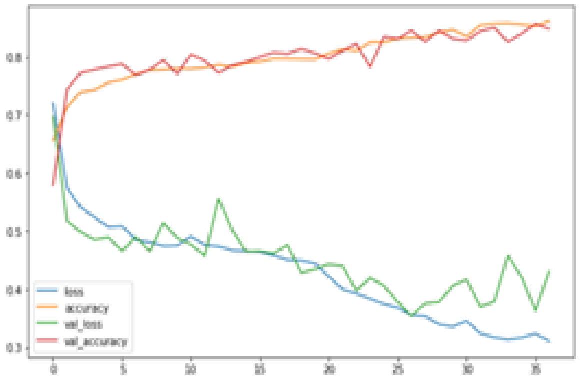

Fig. 7. vRNN loss and accuracy for ELIA 101-1.

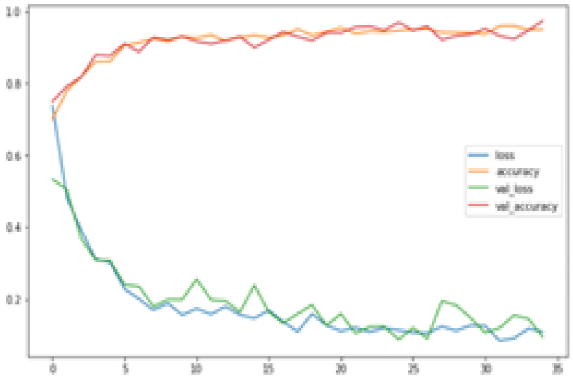

Fig. 8. vRNN loss and accuracy for ELIA 101-2.

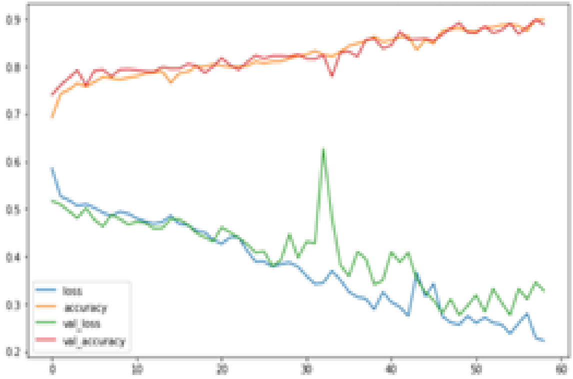

Fig. 9. LSTM loss and accuracy for ELIA 101-1.

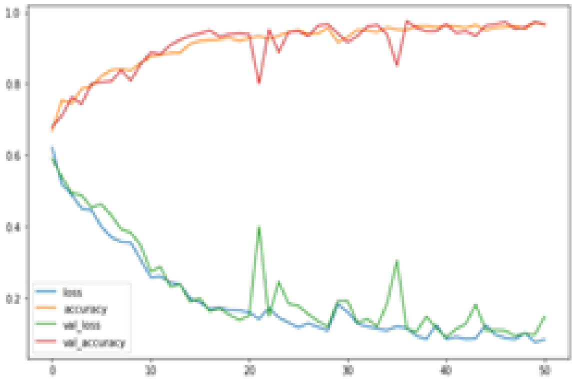

Fig. 10. LSTM loss and accuracy for ELIA 101-2.

with the number of epochs. There is no overfitting since the early stopping was used.

vRNN learned faster than the other models by only 37, and 35 epochs for ELIA 101-1, and ELIA 101-2 datasets, in contrast to the LSTM, which learned by 79, and 51 epochs and GRU, which learned by 88, and 96 epochs, respectively. The slow learning of the LSTM and GRU in comparison with vRNN is due to the models' complexity and the extra wights. This begs the question of the feasibility of using the GRU model on huge students' real-time datasets and the expected time-consuming performance on such data, while considering

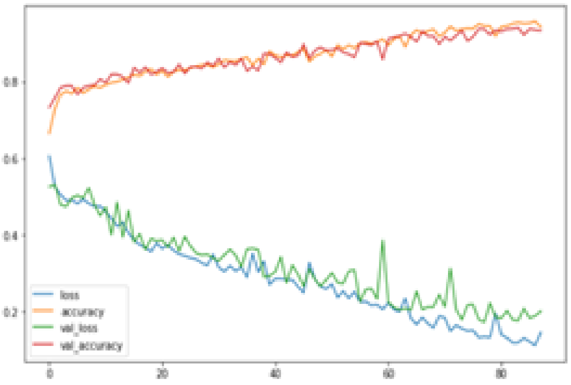

Fig. 11. GRU loss and accuracy for ELIA 101-1.

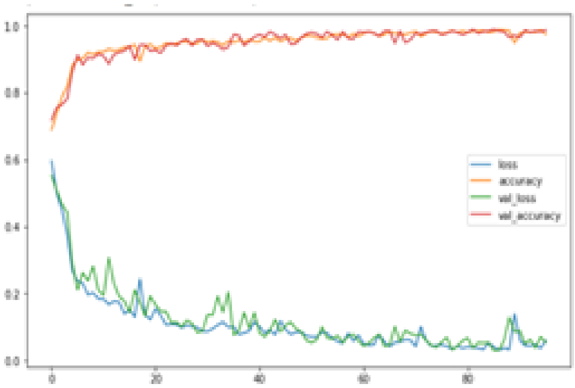

Fig. 12. GRU loss and accuracy for ELIA 101-2.

the sensitivity of at-risk students' activity.

The main objective of this research has been the utilization of DL techniques and models to predict at-risk students, who are, in this learning scenario, the failing students. Therefore, it is of importance that we reflect on the results of the prediction accuracy when it comes to identifying at-risk students. The confusion matrices in Fig. 13, 14, 15, 16, 17, 18, 19, 20 and the Precision and Recall values in Table IV, have demonstrated that the GRU's Recall (representing in our case the failing students) on the (ELIA 101-2) dataset (=97.7%) is the highest value indicative of the model's performance in regard to detecting the at-risk factor, which is critical for our study. The vRNN Recall on the (ELIA 101-2) is also high with a value of 93.8%, which is again representative of the model's capability at classifying the students' fail instances. The GRU's accuracy in identifying failing students on the (ELIA 101-1) comes next with a value of 89.22%, then the LSTM's Recall value (=88.5%) on the same dataset.

The previously cited results point out the predictive power of the GRU, vRNN, and LSTM, almost in that order, especially with reference to detecting at-risk students, on both datasets with Recall values ranging from 81.68% to 97.79%. Though the results in this regard are relatively close, yet the slight distinctions are probably suggestive of two main things. Firstly, there is a differentiation of learning patterns among at-risk students, who are subject to various life situations that impact their learning interactions online which could explain the distinction in the models' performance regarding the detection of their Fail status. Secondly, there is also no uniformity among at-risk students enrolled in different, time-displaced cohorts with regards to their level of expected learning interactions and the impact of those interactions on their academic performance.

It is also significant to reflect on the models' performance with reference to its ability to differentiate passing from failing students. As the confusion matrices in Fig. 13, 14, 15, 16, 19, 20, 17, and 18 show, except of course for the MLP implementation on the (ELIA 101-1) dataset, there is a greater reported accuracy when it comes to predicting passing students in contrast to failing ones. For the GRU, for example, the model has been able to predict 97.8% of passing students and 89.2% of the failing ones on the (ELIA 101-1) dataset in comparison to 100% accuracy of predicting passing students and 97.7% of failing students on the (ELIA 101-2) dataset. As far as the LSTM and vRNN models are concerned, we notice the same trend of the model's ability to predict successful students with a pass indicator. Whereas, the vRNN application to the first dataset, for instance, has resulted in an accurate prediction of 86.5% passing students and 81.6% of the ones with a failing status, its implementation on the second dataset has yielded a 100% accuracy value for predicting passing students and a 93.8% for students with a failure grade. Almost the same results are reported for the LSTM, with accuracy values for predicting successful learners (97.6%, 92.7%) that are higher than the values of classifying unsuccessful ones (88.5%, 85.1%) on both datasets.

This could be attributed to the almost inherent homogenous nature of intermediate to high-performing students' learning activity and behavior which is often intrinsically motivated to the point that their commitment to the learning process manifests in the form of almost uniform patterns across most learning cohorts. DL models, including the ones developed for this study, discover the hidden patterns in learners' data that contribute to an understanding of their learning behaviour, and this reflects positively on the performance of the models.

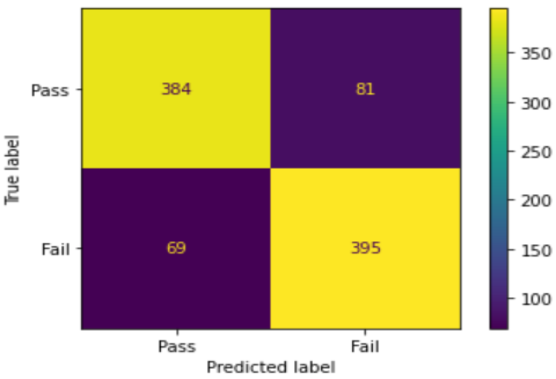

Fig. 13. MLP confusion matrix for ELIA 101-1.

One has to note that the improved performance of the models on the ELIA 101-2 dataset could be attributed to the relatively small number of records originally found in it (1143) including both pass (1137) and fail (6) students, which upon augmentation reached (2274). This, if compared to the

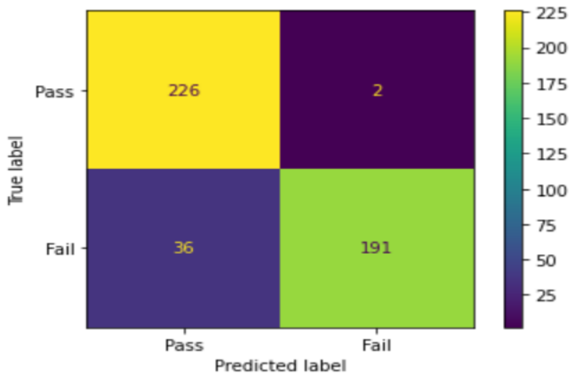

Fig. 14. MLP confusion matrix for ELIA 101-2.

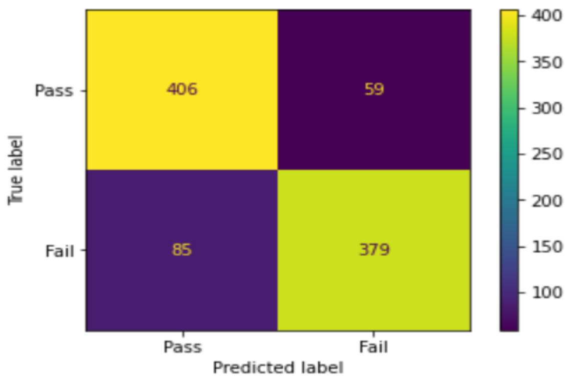

Fig. 15. vRNN confusion matrix for ELIA 101-1.

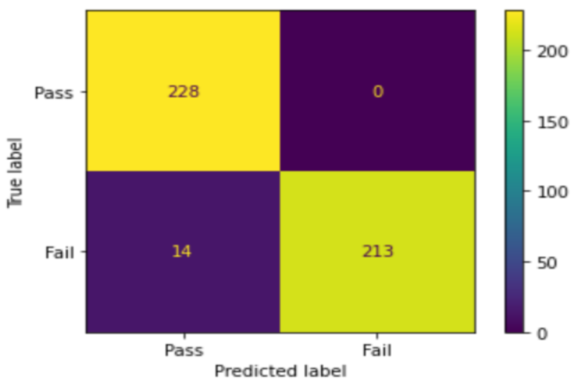

Fig. 16. vRNN confusion matrix for ELIA 101-2.

4644 augmented records in the ELIA 101-1 dataset can be considered one of the limitations decision-makers encounter in real-life educational scenarios. Unlike the MOOC, for example, learning experience, students in compulsory university education are usually grouped in smaller cohorts and required to complete courses within a specific timeframe. Rarely, if ever, especially in Foundation courses, we encounter high rates of failure or even dropout.

Another limitation we believe is related to the dataset size and representation of one course and student cohort. More testing on differentiated datasets representing students' LMS behaviour in KAU can lead to a better understanding of what constitutes risk factors for students.

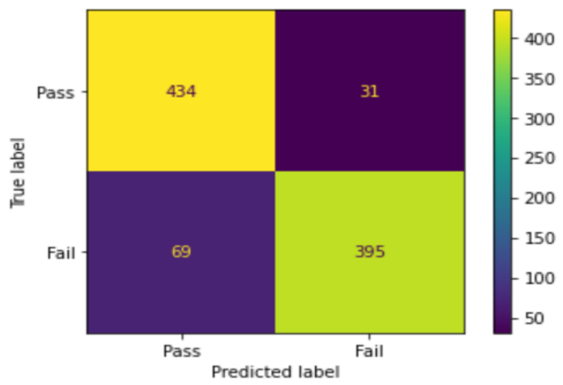

Fig. 17. LSTM confusion matrix for ELIA 101-1.

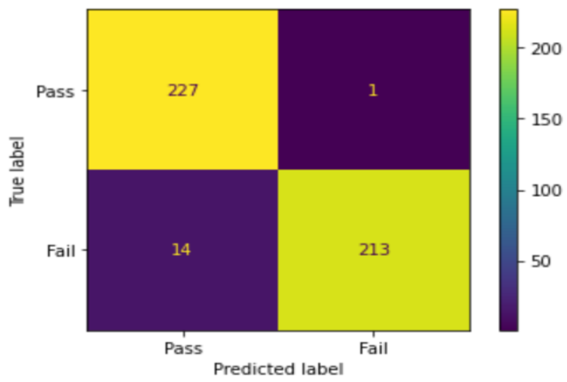

Fig. 18. LSTM confusion matrix for ELIA 101-2.

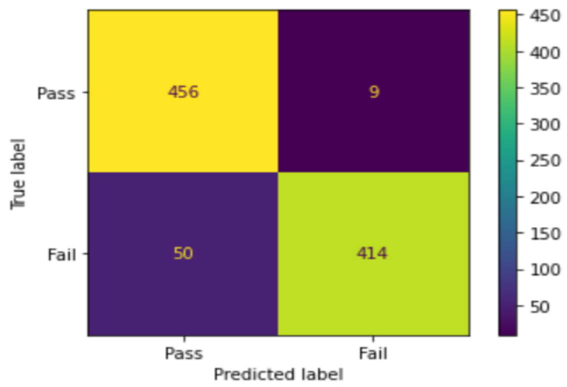

Fig. 19. GRU confusion matrix for ELIA 101-1.

V. Conclusion and Future Work

LMS platforms provide useful information about students' interactions, which can be used to identify at-risk students. In this study, we proposed three neural network models (vRNN, LSTM, and GRU) for predicting both students' final grade performance and at-risk standing based on two datasets extracted from the A4L: KAU Blackboard.

The results show that the GRU performs better than

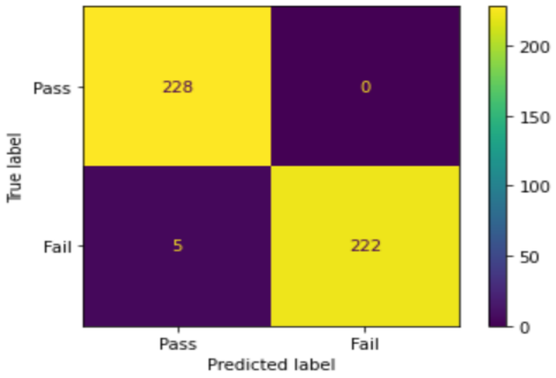

Fig. 20. GRU confusion matrix for ELIA 101-2.

other models in detecting learners' Pass/Fail status because it achieved the highest accuracy 93.65 (on the ELIA 101-1) and 98.90 (on the ELIA 101-2).

As far as predicting the at-risk students (likely to Fail), the previously mentioned results highlight the predictive power of the GRU, vRNN, and LSTM, respectively, on both datasets, with Recall values ranging from 81.68% to 97.79%.

The researchers think the dataset's size and its representation of only one course and student cohort are a drawback. A deeper knowledge of what comprises risk variables for students may result from more testing on differentiated datasets representing students' LMS behaviour in KAU.

For further research, we will use methods to overcome the impact of small size datasets on the realistic performance of DL models by, for example, through implementing advanced data augmentation techniques, considering time-series factors to predict at-risk students half-away through the semester; and, adding other predicators of students' user behaviour inside the LMS and exploring their relation to students' final achievement. More importantly, and while observing the variation among the proposed models in the accuracy of predicting at-risk students on different datasets, we will experiment with ensemble techniques, where the best results of each model might be enhanced by its combination with the others.

Acknowledgment

The researchers would like to thank the Deanship of E-Learning and Distance Education at King Abdulaziz University for the provision of datasets.

References

[1] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches to educational data mining," *Complexity*, vol. 2019, 2019.

[2] L. M. Nkomo and M. Nat, "Student engagement patterns in a blended learning environment: an educational data mining approach," *TechTrends*, vol. 65, no. 5, pp. 808–817, 2021.

[3] Y. Salal, S. Abdullaev, and M. Kumar, "Educational data mining: Student performance prediction in academic," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 4C, pp. 54–59, 2019.

[4] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," *Computers in Human behavior*, vol. 104, p. 106189, 2020.

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

[6] P. Ibañez, C. Villalonga, and L. Nuere, "Exploring student activity with learning analytics in the digital environments of the nebrija university," *Technology, Knowledge and Learning*, vol. 25, no. 4, pp. 769–787, 2020.

[7] Y. He, R. Chen, X. Li, C. Hao, S. Liu, G. Zhang, and B. Jiang, "Online at-risk student identification using rnn-gru joint neural networks," *Information*, vol. 11, no. 10, p. 474, 2020.

[8] J. Sanders, R. Munford, and J. Boden, "Improving educational outcomes for at-risk students," *British Educational Research Journal*, vol. 44, no. 5, pp. 763–780, 2018.

[9] A. Alhassan, B. Zafar, and A. Mueen, "Predict students' academic performance based on their assessment grades and online activity data," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 4, 2020.

[10] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, "Detecting at-risk students with early interventions using machine learning techniques," *IEEE Access*, vol. 7, pp. 149 464–149 478, 2019.

[11] L. A. Buschetto Macarini, C. Cechinel, M. F. Batista Machado, V. Faria Culmant Ramos, and R. Munoz, "Predicting students success in blended learning—evaluating different interactions inside learning management systems," *Applied Sciences*, vol. 9, no. 24, p. 5523, 2019.

[12] P. Kumari, P. K. Jain, and R. Pamula, "An efficient use of ensemble methods to predict students academic performance," in *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*. IEEE, 2018, pp. 1–6.

[13] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, "Towards developing hybrid educational data mining model (hedm) for efficient and accurate student performance evaluation," *Soft Computing*, vol. 24, no. 24, pp. 18 477–18 487, 2020.

[14] E. Howard, M. Meehan, and A. Parnell, "Contrasting prediction methods for early warning systems at undergraduate level," *The Internet and Higher Education*, vol. 37, pp. 66–75, 2018.

[15] J.-L. Hung, K. Rice, J. Kepka, and J. Yang, "Improving predictive power through deep learning analysis of k-12 online student behaviors and discussion board content," *Information Discovery and Delivery*, 2020.

[16] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting students' performance using machine learning techniques," *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, vol. 27, no. 1, pp. 194–205, 2019.

[17] S. Sultana, S. Khan, and M. A. Abbas, "Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts," *International Journal of Electrical Engineering Education*, vol. 54, no. 2, pp. 105–118, 2017.

[18] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," *Education and Information Technologies*, vol. 25, no. 3, pp. 1913–1927, 2020.

[19] S. Hussain, Z. F. Muhsion, Y. K. Salal, P. Theodorou, F. Kurtoglu, and G. Hazarika, "Prediction model on student performance based on internal assessment using deep learning." *iJET*, vol. 14, no. 8, pp. 4–22, 2019.

[20] ELI, "English Language Institute - Preparatory Year Program — eli.kau.edu.sa," https://eli.kau.edu.sa/Pages-preparatory-year-program-en.aspx, 2021.

[21] C. P. Chai, "The importance of data cleaning: Three visualization examples," *Chance*, vol. 33, no. 1, pp. 4–9, 2020.

[22] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[23] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *Int. J. Recent Trends Eng. Res*, vol. 3, no. 4, pp. 444–449, 2017.

[24] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[25] X. Ying, “An overview of overfitting and its solutions,” in *Journal of Physics: Conference Series*, vol. 1168, no. 2. IOP Publishing, 2019, p. 022022.

[26] M. Camacho Olmedo, M. Paegelow, J.-F. Mas, and F. Escobar, “Multilayer perceptron (mlp). geomatic approaches for modeling land change scenarios. an introduction,” in *Geomatic Approaches for Modeling Land Change Scenarios*. Springer, 2018, pp. 1–8.

[27] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.