Toward Better Grade Prediction via A2GP - An Academic Achievement Inspired Predictive Model

Wei Qiu
Nanyang Technological University
qiuwei@ntu.edu.sg

S. Supraja
Nanyang Technological University
supraja.s@ntu.edu.sg

Andy W. H. Khong
Nanyang Technological University
andykhong@ntu.edu.sg

ABSTRACT

Predicting student performance in an academic institution is important for detecting at-risk students and administering early-intervention strategies. We propose a new grade prediction model that considers three factors: temporal dynamics of prior courses across previous semesters, short-term performance consistency, and relative performance against peers. The proposed architecture comprises modules that incorporate the attention mechanism, a new short-term gated long short-term memory network, and a graph convolutional network to address limitations of existing works that fail to consider the above factors jointly. A weighted fusion layer is used to fuse learned representations of the above three modules—course importance, performance consistency, and relative performance. The aggregated representations are then used for grade prediction which, in turn, is used to classify at-risk students. Experiment results using three datasets obtained from over twenty thousand students across seventeen undergraduate courses show that the proposed model achieves low prediction errors and high F1 scores compared to existing models that predict grades and thereafter identifies at-risk students via a pre-defined threshold.

Keywords

Grade prediction, machine learning, attention mechanism, long short-term memory network, graph convolutional network

1. INTRODUCTION

Learning analytics involves the process of collecting, analyzing, and reporting of data generated by learners in an education setting. It optimizes learning and the environment by gaining insights into the learning behavior and/or learner achievements [39]. Among the several sub-disciplines that learning analytics transcends across, prediction of academic performance has received increasing attention in recent years and remains one of the most challenging tasks. Grade prediction plays a central role in the development of data-informed approaches for early-intervention strategies and it is therefore important to achieve a low prediction error—errors leading to high false alarms will result in reduced morale and inefficient allocation of resources while missed detection often results in sustained poor performance [29]. After grade prediction, at-risk students are identified as those whose performance satisfy a pre-defined set of conditions (e.g., those who score below the passing mark in one or more courses).

W. Qiu, S. Supraja, and A. W. H. Khong. Toward better grade prediction via A2GP - an academic achievement inspired predictive model. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 195–205, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6852984

1.1 Existing Works for Grade Prediction

Prediction of academic performance in the form of grade point averages [15], examination grades [13], or academic achievements [25] can be achieved via a variety of sources. These sources include (but not limited to) online learning activities [22, 28, 44], co-curricular activity records [8], demographics [38], and course grades obtained from previous semesters [13,23,32,34]. While online learning offers numerous opportunities for the exploitation of data associated with learning behaviors (in the form of clickstreams and/or online assessment results) [46], many academic institutions still rely on face-to-face instructions for some courses. Extraction of learning behaviors for these courses via audio/visual capturing devices may present challenges in terms of technological capability and privacy concerns. In addition, co-curricular activity records and demographic profiles may not be readily available due to the general personal data protection policies [3]. Therefore, grade prediction using past examination records as useful features [45] has been the main focus in recent years since examination results are often made accessible to policy makers, administrators, instructors, and student care support personnel involved in developing and administering intervention strategies.

Machine learning techniques have been proposed to predict grades of a given (pilot) course based on those achieved in historical (prior) courses. These models exploit the temporal dynamics of student performance across semesters from two aspects—consistency in academic performance and course importance [47]. These aspects have been modeled using a sequential model and the attention mechanism [24], respectively. Sequential models such as the long short-term memory (LSTM) has been applied to model long-term dependencies in online interaction [17] and to predict the grade point average of a given semester from marks obtained across various courses [33]. More recently, as opposed to predicting the aggregated performance for a semester, the LSTM was trained to predict the grade of each course [13]. In

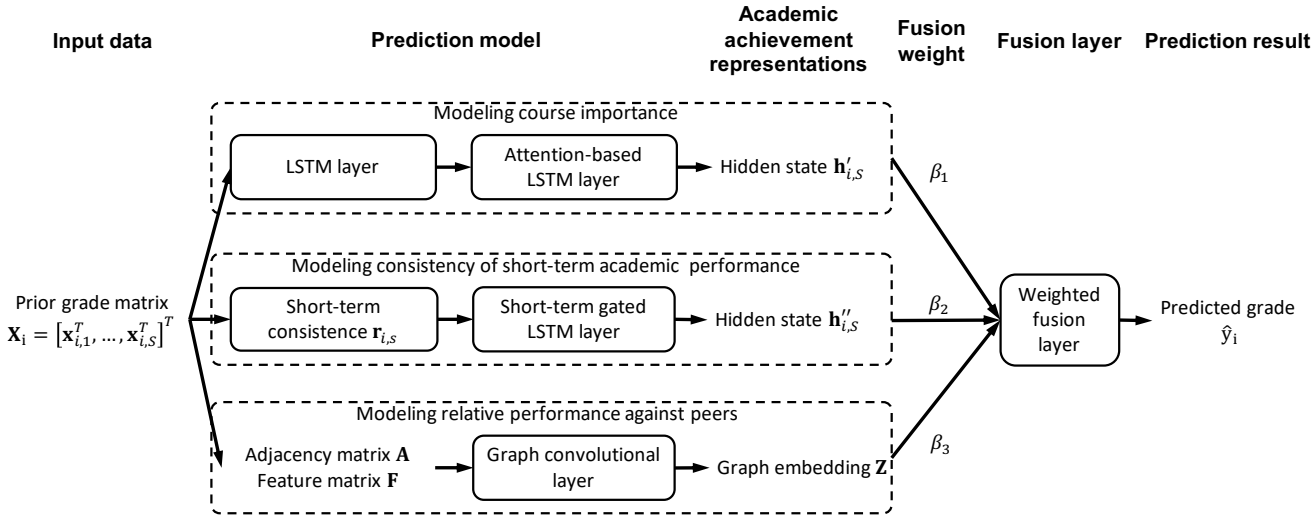

Figure 1: The proposed academic achievement-based grade prediction (A2GP) architecture.

this model, a vector representation of course grades achieved in each of the previous semester was used as input for the LSTM model.

Notwithstanding the above, a course-specific regression model that predicts the grade of a pilot course as a sparse linear combination of prior grades has been proposed [35]. Results presented highlight the detrimental effects of missing regressors for students who have yet attempted an "important" prior course. While modeling of the temporal dynamics of student performance along with the incorporation of the attention mechanism for grade comparison between students has been proposed [30], the intrinsic formulation of LSTM limits its ability to model short-term consistency [21]. Despite the use of knowledge distillation [27], such models do not consider peer performance among students.

In recent years, graph convolutional network (GCN) have been employed to generate meaningful feature representations. In contrast to the use of grade vectors in temporal modeling approaches, these representations model the transitions of grade distributions between courses across semesters [12]. Here, the performance of each student is considered for courses taken consecutively. More recently, nodes representing either students or courses have been used to construct student-course, student-student, and course-course graphs [23]. These graphs consist of edge links computed via grade distribution similarities; they do not model both the long- and short-term sequential information of each student.

While the above techniques achieve good prediction performance, the models are optimized independently and do not consider all the above-mentioned aspects jointly. A holistic approach toward predicting academic performance is important and motivated, in part, by Walberg's theory of educational productivity. Apart from external variables such as quality of instruction and climate, student-centric variables that include prior achievement and student cognitive capacity will influence the academic performance of an individual [43].

1.2 The Proposed Model Architecture

Inspired by student-centric factors highlighted in Section 1.1, we propose an academic achievement-based grade prediction (A2GP) architecture that jointly models the (i) importance of prior courses, (ii) short-term consistency in academic performance across previous semesters, and (iii) benchmarking of student performance relative to their peers. With reference to Figure 1, the first module of the proposed architecture comprises an attention-based LSTM network that encodes the influence of prior course grades on the pilot course. This module is based on existing sequential models (such as LSTM) that have been employed to capture the temporal dynamics of past academic performance [13]. These models are motivated by studies that have established the association between course orderings and academic performance [9, 26]. Such an association is not surprising given that the constructivist approach has often been adopted for curriculum design, resulting in the influence of various prior courses on a pilot course [37]. Such a constructivist strategy has also shown to be effective in terms of academic achievement [2, 18] and improving content mastery that requires higher cognitive levels [1]. Temporal modeling of performance using LSTM is also in line with the Tinto's Student Integration Model which posits that persistence in higher education is a temporal process [5], i.e., the ability to achieve learning outcomes of fundamental courses will influence that of other advanced or related courses. This is further justified if the prior course serves as a pre-requisite for the pilot course.

Compared to existing models that model long-term dynamics of academic achievements, the second module consists of a new short-term gated LSTM (STG-LSTM) that models short-term consistency in academic performance for each student. This module is motivated by the need to consider academic momentum that highlights the influence of workload (which varies across courses and semesters) on academic performance and the achievement of learning out-

comes [14]. Short-term consistency may also arise from academic performance being highly dependent on multiple (yet often convoluted) factors such as socio-economic, psychological, and environmental conditions that a student may face in the recent (past) semesters [36]. From developmental perspective, knowledge inquiry is known to evolve over a series of micro-development resulting in short-term variations in performance [7]. This also aligns with findings that demonstrate the positive effect of mastery and performance goals on short-term and long-term consequences of student achievement [10]. The identification of these patterns would therefore lead to more effective grade prediction.

Beyond representing the performance of a student over time, students are often deemed as at-risk if their performance is consistently below par compared to their peers. In particular, for courses perceived as easy which most students achieve a high grade, achieving a reasonable grade (e.g., Grade C) may still constitute as at-risk when most peers achieved a Grade A. Conversely, a Grade C may not be inferred as being at-risk when most peers achieved similar (or lower) grades for a course perceived by most as challenging. Accounting for such benchmarking of grades is important since such relative performance has shown to achieve lower grade prediction bias than one based on absolute grades [4, 41, 42]. In light of the above, the third module involves a graph convolutional network that models grade differences between student pairs across all prior courses taken by them. This module exploits information derived from students who perform similarly/dissimilarly across commonly taken courses and models such representation that describes the relative performance between students.

For grade prediction, learned academic achievement representations associated with temporal dynamics of past performance, short-term performance consistency of an individual, and relative performance against peers are synthesized via a weighted fusion layer. We formulate a learnable parameter in this layer that determines the relative emphasis of factors influencing the pilot course grade. Learning the weightings for these academic achievement representations is important to model the underlying characteristics of a dataset that contribute to the joint optimization of the model. Performance of the proposed A2GP architecture is evaluated over three student performance datasets obtained over seventeen courses from over twenty thousand students in a university. Results obtained highlighted that the A2GP model improves the performance of LSTM and GCN by 19.0% and 63.3%, respectively, in terms of F1 score for at-risk classification.

This paper is organized as follows: the problem statement and background formulations are described in Section 2. Technicalities of the proposed A2GP model are detailed in Section 3. Details of the datasets, as well as, the comparison analysis with discussions are described in Section 4 while Section 5 concludes the paper.

2. PRELIMINARIES

2.1 Problem Statement

The problem of grade prediction and at-risk detection from prior grades can be described by defining, for each student index i, the exam grade $x_{i,l,s}$ achieved for a prior course l during semester s. Denoting L as the total number of prior courses, a $1 \times L$ grade vector for semester s is given by

$$\mathbf{x}_{i,s} = \left[x_{i,1,s}, \ldots, x_{i,L,s}\right], \tag{1}$$

where $x_{i,l,s} = \phi$ is a null element corresponding to an unregistered course l in that semester. The prior course grades of a student across S number of semesters under consideration can then be represented as an $L \times S$ matrix

$$\mathbf{X}_i = \left[\mathbf{x}_{i,1}^T, \ldots, \mathbf{x}_{i,S}^T\right]. \tag{2}$$

With the above, columns of $\mathbf{X}_i$ form a sequence of vectors that encapsulates the ability of a student to achieve learning outcomes (measured by grades). Given the database of (student, course, grade) up to semester S, the aim is to predict grades for each student on courses he/she will be enrolling in the coming semester.

2.2 Modeling Long-term Dynamics of Academic Performance using LSTM

To model the academic performance across semesters, $\mathbf{X}_i$ serves as features for the prediction of course grades in the forthcoming examinations [13]. The use of $\mathbf{X}_i$, therefore, allows the model to account for courses that a student has re-attempted. An LSTM unit in semester s is described by

$$\mathbf{h}_{i,s} = \text{LSTM}(\mathbf{x}_{i,s}, \mathbf{h}_{i,s-1}), \tag{3}$$

where the above compact form is defined by

$$\mathbf{f}_{i,s} = \sigma(W_f \cdot \mathbf{x}_{i,s} + V_f \cdot \mathbf{h}_{i,s-1} + \mathbf{b}_f), \tag{4a}$$
$$\mathbf{u}_{i,s} = \sigma(W_u \cdot \mathbf{x}_{i,s} + V_u \cdot \mathbf{h}_{i,s-1} + \mathbf{b}_u), \tag{4b}$$
$$\mathbf{o}_{i,s} = \sigma(W_o \cdot \mathbf{x}_{i,s} + V_o \cdot \mathbf{h}_{i,s-1} + \mathbf{b}_o), \tag{4c}$$
$$\widetilde{\mathbf{c}}_{i,s} = \sigma(W_c \cdot \mathbf{x}_{i,s} + V_c \cdot \mathbf{h}_{i,s-1} + \mathbf{b}_c), \tag{4d}$$
$$\mathbf{c}_{i,s} = \mathbf{f}_{i,s} \odot \mathbf{c}_{i,s-1} + \mathbf{u}_{i,s} \odot \widetilde{\mathbf{c}}_{i,s}, \tag{4e}$$
$$\mathbf{h}_{i,s} = \mathbf{o}_{i,s} \odot \tanh(\mathbf{c}_{i,s}). \tag{4f}$$

The variables $\mathbf{u}$, $\mathbf{f}$, $\mathbf{o}$ and their associated subscripts denote the update, forget, and output gates, respectively. These control gates regulate information to be stored in the cell state $\mathbf{c}_{i,s}$ in (4e) allowing LSTM to achieve long-term memory [11]. The weight matrix for input $\mathbf{x}_{i,s}$ and hidden state $\mathbf{h}_{i,s-1}$ in different gate units are denoted, respectively, by matrices W and V. The variables $\widetilde{\mathbf{c}}_{i,s}$ and $\mathbf{b}$ are defined, respectively, as as the cell input activation vector and bias while σ and tanh are the activation functions. The symbol $\odot$ denotes element-wise multiplication. The ability of a student to achieve the course learning outcomes is therefore encoded in hidden states $\mathbf{h}_{i,s}$, which are then updated when grades for the new courses are made available. Predicted grade $\widehat{y}_i$ is then achieved via a fully-connected (FC) layer

$$\widehat{y}_i = \mathbf{w} \cdot \mathbf{h}_{i,S}^T + b, \tag{5}$$

where $\mathbf{w}$ and b are defined, respectively, as the weight vector and bias scalar for the FC layer.

2.3 Modeling Relative Performance using GCN

GCN has been applied to model the interactions between nodes in a graph network. As opposed to [12], where the GCN models transitions between courses across semesters, we define $\mathbf{A}$ as the adjacency matrix such that its elements

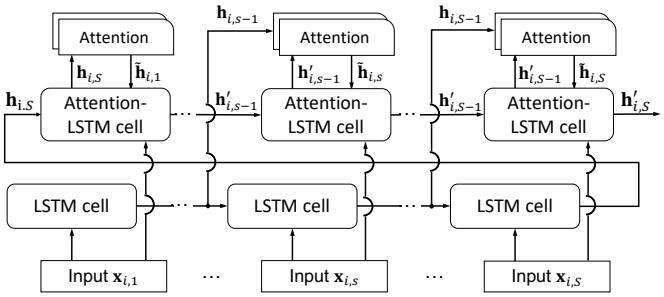

Figure 2: Modeling course importance.

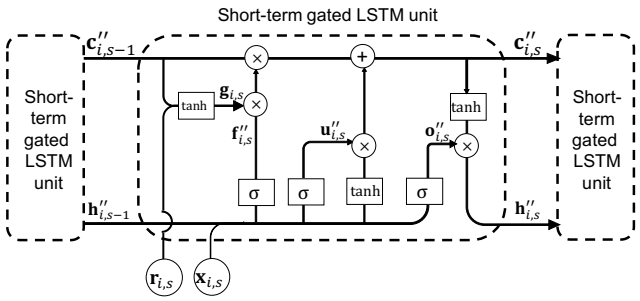

Figure 3: Modeling consistency in student short-term performance.

$A_{i,j}$ denotes the similarity of prior grades between two students—a value of 1 is assigned between students i and j having exactly the same prior grades. The model incorporates a feature matrix $\mathbf{F}$ with elements corresponding to first attempt grades that a student obtained over the past semesters for each prior course. Multiple layers of GCNs are then applied to $\mathbf{A}$ and $\mathbf{F}$ with the $(g+1)$th layer being computed via [20]

$$\mathbf{Z}^{(g+1)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{Z}^{(g)}\mathbf{W}^{(g)}\right). \tag{6}$$

Here, $\mathbf{D}$ is the normalization matrix, $\mathbf{Z}^{(g)}$ is the input to the next layer, and $\mathbf{W}^{(g)}$ is the weight matrix. With the input of the first GCN layer being $\mathbf{Z}^{(0)} = \mathbf{F}$, the output of the last GCN layer is the student-specific graph embedding matrix

$$\mathbf{Z}^{(G)} = \left[\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T\right]^T, \tag{7}$$

where G is the number of GCN layers. Each node embedding (row) vector $\mathbf{z}_i$ for each node (student) then serves as an input to the subsequent FC layer for grade prediction.

3. THE PROPOSED A2GP MODEL

3.1 Modeling Course Importance

Inspired by the attention mechanism in sequence-to-sequence models [24], the first module comprises an LSTM layer and the attention-LSTM layer to model course importance. To formulate the above and as shown in Figure 2, columns of the grade matrix $\mathbf{X}_i$ in (2) serve as input sequences to the LSTM and the ability to achieve the course learning outcomes is therefore encoded in hidden states $\mathbf{h}_{i,s}$ defined in (4f).

The last hidden state $\mathbf{h}_{i,S}$ will be used to initialize the hidden state of the subsequent attention-LSTM layer. This layer is necessary to account for the influence of the various prior courses on the pilot course of interest. The attention mechanism in this layer can be described by first defining

$$\widetilde{\mathbf{h}}_{i,s} = \begin{cases} \mathbf{h}_{i,S}, & \text{if } s = 1; \\ \mathbf{W}_h \cdot \left[\alpha_{i,s-1} \odot \mathbf{h}_{i,s-1}; \mathbf{h}'_{i,s-1}\right]^T, & \text{if } s > 1 \end{cases} \tag{8}$$

as the hidden activation vector, where $\mathbf{h}_{i,s-1}$ is the hidden state of the LSTM layer and $\mathbf{h}'_{i,s-1}$ is the hidden state of the attention-LSTM layer. Here, $\mathbf{W}_h$ is the weight matrix that is to be trained, and the prime notation denotes for the attention layer. Therefore, the hidden activation vector $\widetilde{\mathbf{h}}_{i,s}$ is first initialized as the last hidden state $\mathbf{h}_{i,S}$ from the previous LSTM layer before being updated based on the learned semester-aware attention. Defining $\mathbf{W}_\alpha$ as the weight matrix, the semester-aware attention in (8) for semester s is given by

$$\alpha_{i,s-1} = \text{Softmax}\left(\mathbf{W}_\alpha[\mathbf{h}_{i,s-1}; \mathbf{h}'_{i,s-1}]\right). \tag{9}$$

The hidden state for the next unit $\mathbf{h}'_{i,s}$ (of this attention layer) is then computed using student prior grade $\mathbf{x}_{i,s}$ and the hidden activation vector $\widetilde{\mathbf{h}}_{i,s}$ such that

$$\mathbf{h}'_{i,s} = \text{Attention-LSTM}(\mathbf{x}_{i,s}, \widetilde{\mathbf{h}}_{i,s}), \tag{10}$$

where the Attention-LSTM$(\mathbf{x}_{i,s}, \widetilde{\mathbf{h}}_{i,s})$ is defined in the same form of LSTM$(\mathbf{x}_{i,s}, \mathbf{h}_{i,s})$ in (3) except for the additional attention and hidden activation vector computed in (8) and (9).

We note from (8) that $\widetilde{\mathbf{h}}_{i,1}$ is generated from $\mathbf{h}_{i,S}$, which encodes the academic performance over past semesters. This allows the attention-LSTM layer to incorporate both aggregated and semester-based information simultaneously when computing the semester-aware attention for each semester. The last hidden state $\mathbf{h}'_{i,S}$ of the attention-LSTM layer is then used along with short-term performance consistency and the relative performance representation (described below) for the grade prediction.

3.2 Modeling Consistency in Short-term Performance

Modeling short-term variations in academic performance is necessary since such variations may result from active (or the lack of) intervention strategies administered by the academic institution or changes in social-economic status that distracts students away from their academic pursuit [36]. In the second module, we formulate a short-term gated LSTM that employs, for each semester s, the average examination score computed across three consecutive semesters $s-1$, s and $s+1$, i.e.,

$$\mathbf{r}_{i,s} = \begin{cases} [0, \overline{x}_{i,s}, \overline{x}_{i,s+1}], & \text{if } s = 1; \\ [\overline{x}_{i,s-1}, \overline{x}_{i,s}, \overline{x}_{i,s+1}], & \text{if } 1 < s < S; \\ [\overline{x}_{i,s-1}, \overline{x}_{i,s}, 0], & \text{if } s = S, \end{cases} \tag{11}$$

where $\overline{x}_{i,s}$ is the average of non-empty elements in $\mathbf{x}_{i,s}$ defined in (1). Short-term performance averages are then em-

ployed to update information in the memory cell via

$$\mathbf{g}_{i,s} = \tanh(W_g \cdot \mathbf{r}_{i,s} + V_g \cdot \mathbf{c}_{i,s-1} + \mathbf{b}_g), \tag{12a}$$

$$\mathbf{c}''_{i,s} = \mathbf{g}_{i,s} \odot \mathbf{f}''_{i,s} \odot \mathbf{c}''_{i,s-1} + \mathbf{u}''_{i,s} \odot \widetilde{\mathbf{c}}''_{i,s}, \tag{12b}$$

$$\mathbf{h}''_{i,s} = \mathbf{o}''_{i,s} \odot \tanh(\mathbf{c}''_{i,s}), \tag{12c}$$

where W_g and V_g are the weight matrices and b_g is the bias term for the short-term gate $\mathbf{g}_{i,s}$. The formulations of $\mathbf{f}''_{i,s}$, $\mathbf{u}''_{i,s}$, $\mathbf{o}''_{i,s}$, the input activation function $\widetilde{\mathbf{c}}''_{i,s}$, and the hidden state $\mathbf{h}''_{i,s}$ are identical to those defined in (4). As shown in (12b), both $\mathbf{g}_{i,s}$ and the forget gate $\mathbf{f}''_{i,s}$ control information updates from previous cell state $\mathbf{c}''_{i,s-1}$. The last hidden state $\mathbf{h}''_{i,S}$ is used with other academic achievement representations in the fusion layer for grade prediction.

To gain insights into the above and with reference to Figure 3, the new short-term gate defined in (12a) utilizes $\mathbf{r}_{i,s}$ and $\mathbf{c}''_{i,s-1}$ to determine how short-term consistency in performance affects the cell state update. As opposed to (4a) where the sigmoid function is used, we employ tanh activation in (12a). This is to avoid sharp damp gradients during back propagation, gradient saturation, and gradient updates propagating in different directions when the sigmoid function is used [16, 31]. In addition, $-1 \leq \mathbf{g}_{i,s} \leq 1$ allows the module to model both the positive and negative relationship between performance variation and the cell states.

We also note from (12b) that the previous cell state $\mathbf{c}''_{i,s-1}$ is weighted by both the forget gate $\mathbf{f}_{i,s}$ and the short-term gate $\mathbf{g}_{i,s}$. As per conventional LSTM described by (4a), $\mathbf{f}''_{i,s}$ determines the amount of information to discard from the cell. The short-term gate incorporating $\mathbf{r}_{i,s}$ and $\mathbf{c}''_{i,s-1}$ encapsulates information pertinent to both short-term performance variation and long-term past performance before being passed to the next cell state. Unlike LSTM, where both the input activation $\widetilde{\mathbf{c}}''_{i,s}$ and the previous cell state $\mathbf{c}''_{i,s-1}$ are weighted by a gate learned from the current input and the previous hidden state in (4e), only $\mathbf{c}''_{i,s-1}$ is weighted by $\mathbf{g}_{i,s}$ and $\mathbf{f}''_{i,s}$ in (12b). This is because the input gate $\mathbf{u}''_{i,s}$ does not cater for the removal of information from the cell state. Therefore, weighing the input gate $\mathbf{u}''_{i,s}$ by $\mathbf{g}_{i,s}$ in the second term of (12b) is ineffective. Furthermore, applying such weighting on the output gate $\mathbf{o}''_{i,s}$ may result in relevant information being lost in the next hidden state.

3.3 Modeling Relative Performance Against Peers

The third module models the relative performance between students by employing the graph convolutional layer. We first define a $K \times L$ prior score matrix across all students as

$$\mathbf{F} = \left[\widetilde{\mathbf{x}}_1^T, \ldots, \widetilde{\mathbf{x}}_K^T\right]^T, \tag{13}$$

where K is the total number of students under consideration and $\widetilde{\mathbf{x}}_i = [\widetilde{x}_{i,1}, \ldots, \widetilde{x}_{i,L}]$ with elements $\widetilde{x}_{i,l}$ being the first-attempt grade of the ith student for course l. Unlike (2) where grades across every semester are used in the first module, only first attempts are used for the construction of peer-performance graph since they better represent the ability of the student in achieving the learning outcome compared to his/her peers. In addition, the $\{i, j\}$th element in the proposed $K \times K$ adjacency matrix $\mathbf{A}$ is given by

$$A_{i,j} = \begin{cases} 0, & \text{if } ||\mathbb{N}_{i,j}|| = 0; \\ \rho_{i,j}^{-1} = \left(\frac{\sum_{l \in \mathbb{N}_{i,j}} |\widetilde{x}_{i,l} - \widetilde{x}_{j,l}|}{||\mathbb{N}_{i,j}||}\right)^{-1}, & \text{if } ||\mathbb{N}_{i,j}|| > 0; \\ 1, & \text{if } \widetilde{\mathbf{x}}_i = \widetilde{\mathbf{x}}_j. \end{cases} \tag{14}$$

Here, we define $\mathbb{N}_{i,j}$ as the set of common courses that students i and j have taken and $\|\mathbb{N}_{i,j}\|$ as the number of such courses. Therefore, $\rho_{i,j}$ denotes the average grade difference that students i and j have achieved for these common courses. The above formulation implies that elements of the adjacency matrix $0 \leq A_{i,j} \leq 1$ correspond to the degree of similarity in academic performance between two students.

With $\mathbf{A}$ and $\mathbf{F}$, the GCN encodes peer performance via graph representation $\mathbf{Z}^{(g)}$ computed using (6). We apply two GCN layers and the ith row of $\mathbf{Z}^{(2)}$ (denoted by $\mathbf{z}_i$) constitutes the graph representation corresponding to the relative performance vector for each node (i.e., for the ith student).

3.4 Weighted Fusion Layer and Grade Prediction

To determine the weighting for each academic achievement representation highlighted in Sections 3.1-3.3, a weighted fusion layer is employed. The fusion weight is learned by employing statistics associated with prior and semester average grades. More specifically, we define, for each student i, a 1×4 vector

$$\mathbf{d}_i = \left[\mu_i, \sigma_i, \mu'_i, \sigma'_i\right], \tag{15}$$

where μ_i and σ_i are the mean and standard deviation (STD) of non-empty elements in (2). The variables $\mu'_{i,j}$ and $\sigma'_{i,j}$ denote the mean and STD of

$$\overline{\mathbf{x}}_i = \left[\overline{x}'_{i,1}, \ldots, \overline{x}'_{i,S-1}\right], \tag{16}$$

where $\overline{x}'_{i,s}$ is the average of non-empty elements across two semesters $\mathbf{x}_{i,s-1}$ and $\mathbf{x}_{i,s}$. We note from the above that $\mathbf{d}_i$ incorporates statistical properties associated with both long- and short-term consistency of a student. These features play an important role in influencing the contribution of each academic achievement representation $\mathbf{h}'_{i,S}$, $\mathbf{h}''_{i,S}$, and $\mathbf{z}_i$ to the predicted grade. This dependence is expected given the student prior grade and the short-term consistency have been used as the input to each module to learn the representations.

To determine the weights in the fusion layer, we first define $p = 1, 2, 3$ as the index for the academic achievement representations. Given $\mathbf{d}_i$, these weights are learned via an FC layer given by

$$\beta_{i,p} = \mathbf{w}_p \cdot \mathbf{d}_i^T + b_p, \tag{17}$$

where $\mathbf{w}_p$ and b_p are the trainable weight vector and bias for the pth academic achievement representations. The predicted grade $\widehat{y}_i$ for student i is then given by

$$\widehat{y}_i = \mathbf{w} \cdot \left[\beta_{i,1} \times \mathbf{h}'_{i,S}; \beta_{i,2} \times \mathbf{h}''_{i,S}; \beta_{i,3} \times \mathbf{z}_i\right]^T + b, \tag{18}$$

where $\mathbf{w}$ and b are defined, respectively, as the weight vector and bias scalar for the predictor. We employed the mean-

Algorithm 1 The proposed A2GP architecture

Input: Input sequence $X_i = [\mathbf{x}_{i,1}^T, \ldots, \mathbf{x}_{i,S}^T]$,
Output: Prediction score $\widehat{y}_i$,
Module 1: Modeling course importance:
for student $i \leftarrow 1$ to K **do**
  for $s \leftarrow 1$ to S **do**
    $\mathbf{h}'_{i,s} \leftarrow$ Attention-LSTM$(\mathbf{x}_{i,s}, \widetilde{\mathbf{h}}_{i,s})$ using (10)
  end for
end for
Module 2: Modeling consistency in student short-term performance:
for student $i \leftarrow 1$ to K **do**
  for $s \leftarrow 1$ to S **do**
    $\mathbf{r}_{i,s} \leftarrow (\mathbf{x}_{i,s-1}, \mathbf{x}_{i,s}, \mathbf{x}_{i,s+1})$ using (11)
    $\mathbf{h}''_{i,s} \leftarrow$ Short-term gated LSTM$(\mathbf{x}_{i,s}, \mathbf{r}_{i,s})$ using (12)
  end for
end for
Module 3: Modeling relative performance against peers:
$\mathbf{F} = [\widetilde{\mathbf{x}}_1^T, \ldots, \widetilde{\mathbf{x}}_K^T]^T$
for student $i \leftarrow 1$ to K **do**
  for student $j \leftarrow 1$ to K **do**
    $A_{i,j} \leftarrow (\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_j)$ using (14)
  end for
end for
$Z \leftarrow A, F$ using (6)
Module 4: Weighted-fusion:
for student $i \leftarrow 1$ to K **do**
  $\mathbf{d}_i = [\mu_i, \sigma_i, \mu'_i, \sigma'_i]$
  $\beta_{i,p} = \mathrm{FC}(\mathbf{d}_i)$ using (17)
  $\widehat{y}_i \leftarrow \mathrm{FC}([\beta_{i,1} \times \mathbf{h}'_{i,s}, \beta_{i,2} \times \mathbf{h}''_{i,s}, \beta_{i,3} \times \mathbf{z}_{i,s}])$ using (18)
end for
return $\widehat{y}_i$

square error loss function

$$\mathcal{L} = \frac{1}{K} \sum_{i=1}^{K} (y_i - \widehat{y}_i)^2 \tag{19}$$

to compute the prediction loss for a total of K students. Similar to [13] [33], a student is classified as at-risk if his/her predicted grade for the pilot course is lower than the pre-defined threshold T, i.e., the classification label is computed by

$$\widehat{\varphi}_i = \begin{cases} \text{At-risk}, & \text{if } \widehat{y}_i < T; \\ \text{Non at-risk}, & \text{if } \widehat{y}_i \geq T. \end{cases} \tag{20}$$

In line with Figure 1 that shows the proposed A2GP architecture, Algorithm 1 provides a formal description of the proposed model.

4. RESULTS AND DISCUSSION

4.1 Datasets

Three datasets have been collected from various departments in a local university with institutional review board (IRB) approval that includes the personal data protection policies. Since other open-source datasets employed for grade prediction do not include past snapshots of examination records, performance of the proposed A2GP model and baseline architectures is evaluated on the datasets from only this university. Seventeen courses across these datasets and their corresponding detailed information are illustrated in Table 1. The number of students denoted by "S" and the percentage of at-risk students denoted by "A" in each course for training and testing are also tabulated. These core courses have been offered to all undergraduates across the three engineering departments during their freshman and sophomore years. These datasets include grades obtained by students who are enrolled from academic year (AY) 2015 to 2019. In our context, the pre-defined threshold is $T = 40$ which refers to the passing mark of the university.

Table 1: Number of students in the training and testing set

Department	Course index	Training set		Testing set	
		S	A (%)	S	A (%)
Department 1	$C_{1,1}$	1241	13.54	249	6.83
	$C_{1,2}$	2459	1.75	384	1.04
	$C_{1,3}$	1314	9.97	261	3.83
	$C_{1,4}$	2448	5.35	357	3.92
	$C_{1,5}$	1524	7.68	312	3.85
	$C_{1,6}$	1000	4.70	190	4.74
	Total	9986	6.38	1753	3.76
Department 2	$C_{2,1}$	1001	12.89	223	8.07
	$C_{2,2}$	600	12.33	98	7.14
	$C_{2,3}$	1029	4.76	355	2.25
	$C_{2,4}$	1034	4.84	205	2.44
	$C_{2,5}$	987	6.69	167	3.59
	$C_{2,6}$	1842	7.76	409	3.67
	Total	6493	7.87	1457	4.05
Department 3	$C_{3,1}$	1001	12.89	133	9.02
	$C_{3,2}$	600	12.33	84	8.33
	$C_{3,3}$	1029	4.76	245	3.27
	$C_{3,4}$	1842	7.76	192	5.21
	$C_{3,5}$	165	29.70	32	31.25
	Total	4637	9.58	686	6.85

S: number of students
A: percentage of at-risk students

To reflect real-world deployment, we predict course grades using examination grades obtained in previous semesters. In particular, the model was trained using prior courses from AY2015 to 2018 and to predict grades for courses registered in AY2019. In this work, the prior grade matrix $\mathbf{X}_i$ is of dimension 20×10 representing twenty prior courses and ten semesters across our dataset. While the undergraduate degree program requires eight semesters to complete, ten semesters were included since some students require a longer duration to graduate.

4.2 Implementation and Performance Metrics

The proposed A2GP model is trained using the Adam optimizer [19]. Hyper-parameters for each model were initialized using the Xavier initialization method [6] and the activation function is a rectified linear unit (ReLU). Adopting a systematic approach [40], the learning rate is first initialized with a small value, e.g., 1×10^{-7}, before being increased exponentially to a pre-defined upper-bound, e.g., 10. An optimal learning rate for the model is then determined from the range in which the model experiences the highest rate of decrease in model loss. This corresponds to 0.8×10^{-3} in our experiments.

Performance of the grade prediction models was evaluated via the mean absolute error defined by

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \widehat{y}_i|, \tag{21}$$

which describes the average absolute error between the predicted and target grades for each grade prediction model. Classification performance of at-risk student detection was evaluated using the F1 score

$$F1 = \frac{2PR}{P + R}, \tag{22}$$

which is computed from the recall (R) and precision (P) scores. The recall score quantifies the number of correct at-risk predictions out of all actual at-risk students in the dataset while precision quantifies the number of correct at-risk predictions out of all detected at-risk students.

4.3 Performance Comparison in Terms of MAE and F1

While many grade prediction algorithms exist, we focus on models that rely only on prior grades obtained from previous semesters. This is important in our context since we cannot assume that all courses are offered online (for the extraction of clickstreams) or that we have access to audio-visual information in a physical classroom setting. To this end, we evaluate the proposed A2GP model on each pilot course by comparing its performance with two widely used classification algorithms (logistic regression (LR) and support vector machine (SVM)), and the LSTM [13]. We have also implemented two variants of the LSTM-based model (attn-LSTM [24] and STG-LSTM) and the GCN [20]. The attn-LSTM was implemented using (8)-(10) while the STG-LSTM is defined by (11)-(12). The GCN was modified from [20] by highlighting the relative performance using (14).

The performance of the models for each department in terms of MAE defined by (21) is tabulated in Table 2. With the predicted grades, at-risk classification performance in terms of the F1 score defined by (22) is tabulated in Table 3 for all predictive models under consideration. Results highlight that the proposed A2GP architecture achieves the lowest average MAE and highest average F1 score than all baseline models across all departments. Compared to the variants of LSTM, the LR and SVM models suffer from poor performance for since these two models do not consider the temporal information that is important for grade prediction.

Table 2: Performance comparison of different models using MAE (a lower value indicates better performance)

Methods	Dept. 1	Dept. 2	Dept. 3	Average
LR	0.121	0.137	0.123	0.127
SVM	0.124	0.138	0.121	0.128
LSTM [13]	0.113	0.129	0.111	0.118
attn-LSTM [24]	0.109	0.128	0.120	0.119
STG-LSTM	**0.102**	0.124	0.116	0.114
GCN [20]	0.109	0.121	0.115	0.115
The proposed A2GP model	0.104	**0.119**	**0.109**	**0.111**

Dept.: Department

It is also interesting to note that all models achieve lower grade prediction and at-risk classification performance for Departments 2 and 3 compared to Department 1. This is because Department 1 has a mandatory set of courses for all students while students from Departments 2 and 3 have the freedom to select courses not offered by their respective schools. Due to this difference in the course selection procedure, there are fewer commonly taken courses between students in Departments 2 and 3. Since GCN determines the relative performance of a student in comparison to his/her peers, an insufficient number of overlapping courses makes it more challenging to predict the grades for the pilot courses. On the same note, since there exist various combinations of courses taken by students from Departments 2 and 3, the LSTM-based models (attn-LSTM and STG-LSTM) are unable to identify similar sequences of temporal information among students. Therefore, diversity in the prior courses results in the poor prediction of a common pilot course grade.

In terms of at-risk classification for Department 1, the attn-LSTM model (with an F1 score of 0.324) achieves an approximate 17% improvement compared to LSTM. This improvement is attributed to attn-LSTM being able to detect the prior courses with higher importance compared to LSTM that equally weighs all courses to determine the grade for the pilot courses. The STG-LSTM model, on the other hand, accounts for the short-term fluctuation of student performance during the update of hidden states. This is in contrast to LSTM that updates the hidden state with equal importance applied to previous hidden states irrespective of any variations in performance. This results in a 10% reduction in the average MAE and 30% increase in an average F1 score for STG-LSTM over that of LSTM.

With regard to the non-temporal approach, since the GCN model constructs an input graph based on grade differences between student pairs (for identifying relative performance), variance of grades within a given dataset would determine the extent of distinguishability among the students. Figure 4 shows the relationship between the F1 score obtained using GCN and the standard deviation σ_{C_i} of the prior grades for

Table 3: At-risk prediction performance using F1 score (a higher value signifies better performance)

Department index	Course index	LR	SVM	LSTM [13]	attn-LSTM [24]	STG-LSTM	GCN [20]	The proposed A2GP model
Department 1	$C_{1,1}$	0.372	0.341	0.293	0.4	**0.457**	0.375	0.439
	$C_{1,2}$	0.222	0.222	0.24	0.308	0.4	0.286	**0.444**
	$C_{1,3}$	0.345	0.222	0.348	0.381	0.438	0.308	**0.48**
	$C_{1,4}$	0.174	0.154	0.34	0.218	**0.381**	0.211	0.368
	$C_{1,5}$	0.313	0.294	0.25	0.435	0.467	0.313	**0.526**
	$C_{1,6}$	0.143	0.222	0.143	0.2	0.154	0	**0.286**
	Average	0.261	0.243	0.269	0.324	0.383	0.249	**0.424**
Department 2	$C_{2,1}$	0.291	0.314	0.241	0.321	0.290	0.276	**0.328**
	$C_{2,2}$	0.235	0.264	0.3	0.174	0.25	0.111	**0.348**
	$C_{2,3}$	0.286	0.264	0.318	0.273	0.19	0.231	**0.32**
	$C_{2,4}$	0	0	0.231	0.095	**0.286**	0	0.08
	$C_{2,5}$	0	0.167	0.182	0.222	0.133	0	**0.273**
	$C_{2,6}$	0.313	0.3	**0.341**	0.114	0.217	0.211	0.28
	Average	0.187	0.218	0.269	0.200	0.228	0.138	**0.271**
Department 3	$C_{3,1}$	0.214	0.244	0.267	0.353	0.357	0.308	**0.4**
	$C_{3,2}$	0.3	0.221	**0.353**	0.308	0.174	0.25	0.222
	$C_{3,3}$	0.253	0.244	0.353	0.154	**0.4**	0.333	**0.4**
	$C_{3,4}$	0.1	0.164	**0.235**	0.118	0.077	0	0.167
	$C_{3,5}$	0.221	0.2	0.571	0.615	0.621	0.444	**0.606**
	Average	0.218	0.215	0.356	0.310	0.326	0.267	**0.359**
Average		0.222	0.226	0.295	0.276	0.311	0.215	**0.351**

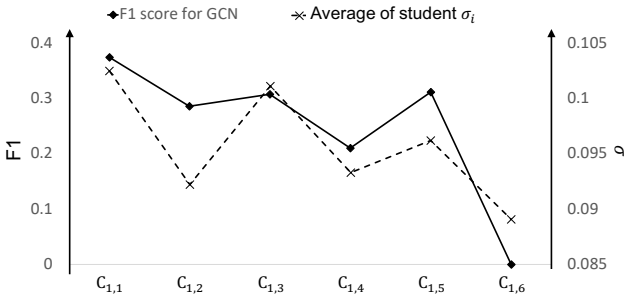

Figure 4: Relationship between F1 score of modeling relative performance against peer and average of σ_i.

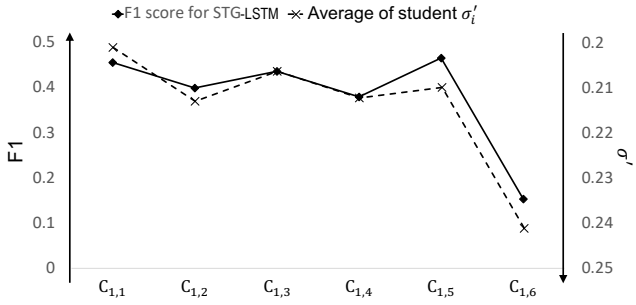

Figure 5: Relationship between F1 score of modeling short-term performance and average of σ_i'.

all students. A smaller value of σ_{C_i} denotes a high similarity in performance among all students taking that course, implying that it is challenging for GCN to differentiate at-risk student performance from well-performing students (low F1 score). Therefore, we note that the ability of GCN for grade prediction is dependent on the underlying statistical properties of a dataset, resulting in varying performance across the courses for Department 1. We also note that GCN is not able to detect any at-risk students for some of the courses (reflected by the F1 score of zero in Table 1) since there were only few students who were actually at-risk. With most of the students achieving good grades in these courses, the performance variation among students is minimal, resulting in GCN not being able to identify relative performance differences.

Compared to the above models, the proposed A2GP model achieves the lowest average MAE of 0.111 and highest average F1 score of 0.351 across the three departments. These results highlight the importance of synthesizing the three dimensions associated with course importance, performance consistency, and benchmarking. Although the performance of individual modules (attention module and short-term gated module) are modestly higher in comparison to LSTM, A2GP includes a weighted fusion that adaptively determines the importance of each module depending on the relevance of the academic achievement representations to each dataset. Figure 6 illustrates the performance of our proposed A2GP model in terms of F1 score with a weighted fusion layer (implemented via (18)) or an equal fusion layer (where β_1= β_2=β_3 = 1). It is important to highlight that the weighted

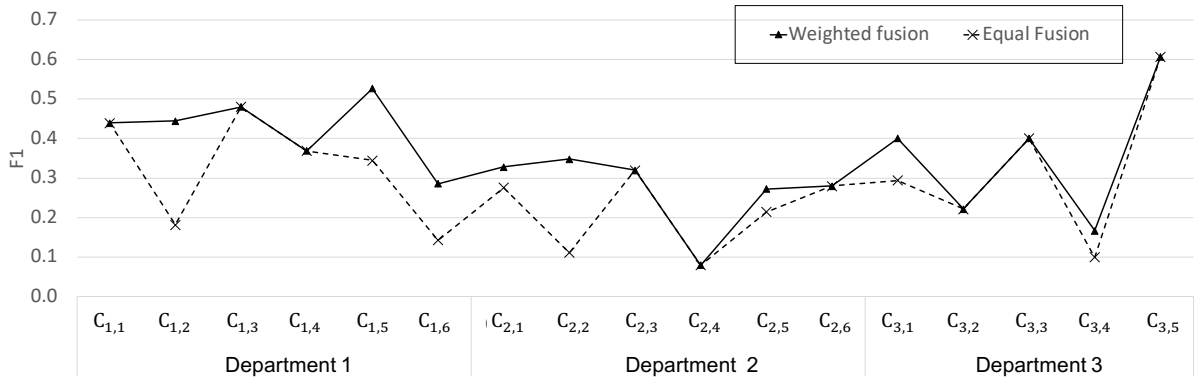

Figure 6: F1 score across courses for model with/without weighted fusion.

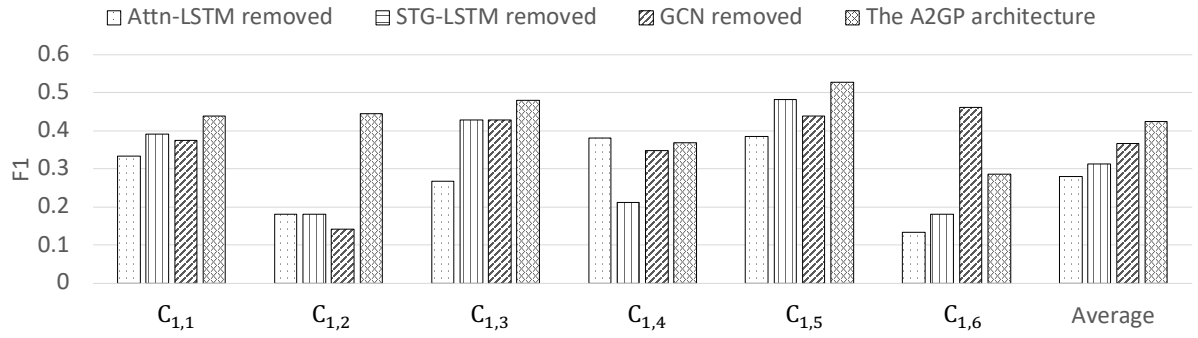

Figure 7: Ablation of the proposed architecture.

fusion layer achieves significantly higher F1 scores for several courses with the remaining courses exhibiting similar performance as that of the equally weighted configuration. In particular, A2GP with weighted fusion (implemented via (17)) achieves an improvement of 0.263 F1 score for course $C_{1,2}$ over that of the equally fusion strategy.

As described in Section 3.4, the underlying statistics of student prior exam grades $\mathbf{d}_i$ have been used to learn the fusion weights $\beta_{i,p}$. Figure 5 shows the variation of F1 with the mean of STD σ' defined by averaging σ'_i according to (15) over all students. Results plotted in this figure was generated by evaluating STG-LSTM over all courses offered by Department 1. A high value of σ' implies many students in this course exhibit short-term performance fluctuations in the past semesters. It can be seen that the F1 score reduces with increasing σ'. This implies that higher fluctuations in short-term student performance will pose a challenge for the model to detect at-risk students. To address limitations faced by individual models, the weighted fusion layer in the A2GP model de-emphasizes the aspects affected by the dataset while emphasizing other academic achievement representations which, in turn, aid the grade prediction process.

4.4 Ablation Test

Figure 7 shows results associated with an ablation test performed on the A2GP model. Here, at-risk prediction performance is determined when each of the academic achievement representation is removed from the A2GP architecture. Among all three representations, the average performance across all courses reduces most significantly when attn-LSTM, i.e., $\mathbf{h}'_{i,s}$ is removed. This is due to the impact of identifying course importance on the entire cohort since constructivist approach is often adopted during curriculum development. The STG-LSTM model, on the other hand, is student-specific—fluctuation in individual performance depends on unique circumstances not generalizable for the other students. Therefore, the A2GP model is less sensitive to STG-LSTM compared to attn-LSTM. Removing GCN results in the least difference in prediction performance since different courses exhibit varying degree of grade spread, with some lacking sufficient information for GCN to discern among student performance resulting in an inappropriate representation. Nonetheless, incorporating peer performance will still be beneficial to A2GP model as seen in Figure 7.

5. CONCLUSIONS AND FUTURE WORK

An academic achievement-based grade prediction architecture is proposed for grade prediction and at-risk student detection. To utilize three important aspects in student prior performance—course importance, short-term performance fluctuation, and relative performance against peers, three modules have been formulated and fused. The first module learns the prior grade representation along with course

importance by employing the attention-based LSTM model. The new STG-LSTM in the second module is motivated by the need to model short-term fluctuation in academic performance. The third module is motivated by the need to model relative performance when detecting at-risk student—students are often deemed as at-risk if their performance is consistently below par compared to their peers. We evaluated the prediction performance of the proposed A2GP model by comparing its performance with baseline models. Results obtained showed that the proposed architecture outperforms existing at-risk detection algorithms across seventeen undergraduate courses from three departments. Improving the F1 score in at-risk student detection facilitates the administration of pre-emptive interventions by instructors, counsellors, or pastoral care managers.

There are possible avenues for future work. First, A2GP has been optimized for use with prior examination grades and has not been validated for online learning activities. Behavioral features associated with the consumption of online assets may provide additional (and often complementary) information that may aid grade prediction. Secondly, further investigations can be performed on the fusion weights defined by (17). These weights may offer insights into the importance of different representations influenced by characteristics that govern the cohort performance within each course.

6. REFERENCES

[1] S. Adak. Effectiveness of constructivist approach on academic achievement in science at secondary level. *Edu. Research Reviews*, 12(22):1074–1079, 2017.

[2] M. F. Ayaz and H. Şekerci. The effects of the constructivist learning approach on student's academic achievement: A meta-analysis study. *Turkish Online J. Edu. Tech.*, 14(4):143–156, 2015.

[3] W. B. Chik. The Singapore personal data protection act and an assessment of future trends in data privacy reform. *Comput. Law & Security Review*, 29(5):554–575, 2013.

[4] R. Dattakumar and R. Jagadeesh. A review of literature on benchmarking. *Benchmarking: Int. J.*, 10(3):176–209, 2003.

[5] E. Fincham, B. Roozemberczki, V. Kovanovic, S. Joksimovic, J. Jovanovic, and D. Gasevic. Persistence and performance in co-enrollment network embeddings: An empirical validation of Tinto's student integration model. *IEEE Trans. Learn. Tech.*, 14(1):106–121, 2021.

[6] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. Int. Conf. Artif. Intell. and Statist.*, pages 249–256, 2010.

[7] N. Granott. How microdevelopment creates macrodevelopment: Reiterated sequences, backward transitions, and the zone of current development. *Microdevelopment: Transition Processes in Development and Learn.*, 7:213–244, 2002.

[8] Q. Guo, M. Chen, D. An, and L. Ye. Prediction of students' course failure based on campus card data. In *Proc. IEEE Int. Conf. Robots & Intell. Sys.*, pages 361–364, 2019.

[9] T. Gutenbrunner, D. D. Leeds, S. Ross, M. Riad-Zaky, and G. M. Weiss. Measuring the academic impact of course sequencing using student grade data. In *Proc. Int. Conf. Educational Data Mining*, pages 1–5, 2021.

[10] J. M. Harackiewicz, K. E. Barron, J. M. Tauer, S. M. Carter, and A. J. Elliot. Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *J. Edu. Psych.*, 92(2):316–330, 2000.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[12] Q. Hu and H. Rangwala. Academic performance estimation with attention-based graph convolutional networks. In *Proc. Int. Conf. Educational Data Mining*, pages 69–78, 2019.

[13] Q. Hu and H. Rangwala. Reliable deep grade prediction with uncertainty estimation. In *Proc. Int. Conf. Learn. Anal. & Knowl.*, pages 76–85, 2019.

[14] N. Huntington-Klein and A. Gill. Semester course load and student performance. *Research in Higher Educ.*, 62(5):623–650, 2021.

[15] Z. Iqbal, A. Qayyum, S. Latif, and J. Qadir. Early student grade prediction: an empirical study. In *Proc. IEEE Int. Conf. Advancements in Computational Sci.*, pages 1–7, 2019.

[16] B. L. Kalman and S. C. Kwasny. Why tanh: choosing a sigmoidal function. In *Proc. IEEE Int. Joint Conf. Neural Netw.*, volume 4, pages 578–581, 1992.

[17] B.-H. Kim, E. Vizitei, and V. Ganapathi. GritNet: Student performance prediction with deep learning. In *Proc. Int. Conf. Educ. Data Mining*, pages 625–629, 2018.

[18] J. S. Kim. The effects of a constructivist teaching approach on student academic achievement, self-concept, and learning strategies. *Asia Pacific Edu. Review*, 6:7–19, 2005.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015.

[20] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proc. Int. Conf. Learn. Representations*, pages 76–85, 2017.

[21] I. Lee, D. Kim, S. Kang, and S. Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1012–1020, 2017.

[22] K. Liu, S. Tatinati, and A. W. Khong. Context-based data model for effective real-time learning analytics. *IEEE Trans. Learn. Technologies*, 13(4):790–803, 2020.

[23] X. Lu, Y. Zhu, Y. Xu, and J. Yu. Learning from multiple dynamic graphs of student and course interactions for student grade predictions. *Neurocomputing*, 431:23–33, 2021.

[24] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proc. Int. Conf. Empirical Methods in Natural Lang. Process.*, pages 1412–1421, 2015.

[25] M. Marschark, D. M. Shaver, K. M. Nagle, and L. A. Newman. Predicting the academic achievement of deaf and hard-of-hearing students from individual, household, communication, and educational factors. *Exceptional Children*, 81(3):350–369, 2015.

[26] H. R. Mkwazu and C. Yan. Grade prediction method for university course selection based on decision tree. In *Proc. Int. Conf. Aviation Safety Info. Tech.*, pages 593–599, 2020.

[27] R. Murata, A. Shimada, and T. Minematsu. Early detection of at-risk students based on knowledge distillation rnn models. In *Proc. Int. Conf. Educational Data Mining*, pages 1–5, 2021.

[28] K. H. R. Ng, S. Tatinati, and A. W. H. Khong. Online education evaluation for signal processing course through student learning pathways. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, pages 6458–6462, 2018.

[29] K. H. R. Ng, S. Tatinati, and A. W. H. Khong. Grade prediction from multi-valued click-stream traces via Bayesian-regularized deep neural networks. *IEEE Trans. Signal Process.*, 69:1477–1491, 2021.

[30] K. Niu, X. Cao, and Y. Yu. Explainable student performance prediction with personalized attention for explaining why a student fails. In *Proc. Int. Assoc. Advancement Artifical Intell. Workshop AI Edu.*, pages 1–11, 2021.

[31] C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall. Activation functions: comparison of trends in practice and research for deep learning. In *Proc. Int. Conf. Computational Sci. and Tech.*, pages 124–133, 2021.

[32] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata. A neural network approach for students' performance prediction. In *Proc. Int. Learn. Anal. & Knowl. Conf.*, pages 598–599, 2017.

[33] A. P. Patil, K. Ganesan, and A. Kanavalli. Effective deep learning model to predict student grade point averages. In *Proc. IEEE Int. Conf. Computational Intell. and Computing Research*, pages 1–6, 2017.

[34] A. Polyzou and G. Karypis. Grade prediction with course and student specific models. In *Proc. Pacific-Asia Conf. Knowl. Discovery and Data Mining*, pages 89–101. Springer, 2016.

[35] A. Polyzou and G. Karypis. Grade prediction with models specific to students and courses. *Int. J. Data Sci. and Anal.*, 2(3):159–171, 2016.

[36] K. Rajandran, T. C. Hee, S. Kanawarthy, L. K. Soon, H. Kamaludin, and D. Khezrimotlagh. Factors affecting first year undergraduate students academic performance. *Scholars J. of Economics, Business, and Management*, 2(1A):54–60, 2015.

[37] S. H. Seyyedrezaie and G. Barani. Constructivism and curriculum development. *J. Humanities Insights*, 1(3):119–124, 2017.

[38] A. M. Shahiri and W. Husain. A review on predicting student's performance using data mining techniques. *Procedia Comput. Sci.*, 72:414–422, 2015.

[39] G. Siemens and D. Gašević. Special issue on learning and knowledge analytics. *Edu. Technol. & Society*, 15(3):1–163, 2012.

[40] L. N. Smith. Cyclical learning rates for training neural networks. In *Proc. IEEE Winter Conf. Appl. of Comput. Vis.*, pages 464–472, 2017.

[41] R. R. Subramanian, D. V. V. S. S. S. Babu, D. U. Rani, M. Devendrareddy, B. Kusuma, and R. R. Sudharsan. Detecting bias in the relative grading system using machine learning. In *Proc. Int. Conf. Advancements Electrical Electronics Comm. Comp. Automation (ICAECA)*, pages 1–6, 2021.

[42] S. Syme, C. Davis, and C. Cook. Benchmarking australian enabling programmes: assuring quality, comparability and transparency. *Assessment Eval. Higher Edu.*, 46(4):572–585, 2020.

[43] H. J. Walberg, B. J. Fraser, and W. W. Welch. A test of a model of educational productivity among senior high school students. *J. Edu. Research*, 79:133–139, 1986.

[44] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. MOOC dropout prediction: How to measure accuracy? In *Proc. ACM Conf. Learn. and Scale*, pages 161–164, 2017.

[45] A. T. Widjaja, L. Wang, T. T. Nghia, A. Gunawan, and E.-P. Lim. Next-term grade prediction: A machine learning approach. In *Proc. Int. Conf. Educational Data Mining*, pages 700–703, 2020.

[46] T. Yang, C. Brinton, C. Joe-Wong, and M. Chiang. Behavior-based grade prediction for MOOCs via time series neural networks. *IEEE J. Sel. Topics in Signal Process.*, 11(5):716–728, 2017.

[47] Y. Zhang, R. An, S. Liu, J. Cui, and X. Shang. Predicting and understanding student learning performance using multi-source sparse attention convolutional neural networks. *IEEE Trans. Big Data*, Early Access:1–14, 2021.