

Deep Learning with Data Transformation and Factor Analysis for Student Performance Prediction

Tran Thanh Dien¹, Sang Hoai Luu², Nguyen Thanh-Hai³, Nguyen Thai-Nghe^{*4}

College of Information and Communication Technology

Can Tho University, Can Tho, Vietnam

* Corresponding Author: Can Tho University, 3/2 Street, Ninh Kieu District, Can Tho city, Vietnam

Abstract—Student performance prediction is one of the most concerning issues in the field of education and training, especially educational data mining. The prediction supports students to select courses and design appropriate study plans for themselves. Moreover, student performance prediction enables lecturers as well as educational managers to indicate what students should be monitored and supported to complete their programs with the best results. These supports can reduce formal warnings and expulsions from universities due to students' poor performance. This study proposes a method to predict student performance using various deep learning techniques. Also, we analyze and present several techniques for data pre-processing (e.g., Quantile Transforms and MinMax Scaler) before fetching them into well-known deep learning models such as Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) to do prediction tasks. Experiments are built on 16 datasets related to numerous different majors with appropriately four million samples collected from the student information system of a Vietnamese multidisciplinary university. Results show that the proposed method provides good prediction results, especially when using data transformation. The results are feasible for applying to practical cases.

Keywords—Deep learning; student performance; mark prediction; Long Short Term Memory (LSTM); Convolutional Neural Networks (CNN); data pre-processing; multidisciplinary university

I. INTRODUCTION

Recently, the number of students who have been warned and forced to leave school tends to increase. One of the reasons can be that students are not able to evaluate and predict correctly his or her ability to select appropriate courses. Student performance is an important task of higher educational institutions because it is a criteria for high quality universities that are based on excellent profile of their academic achievements. There are several definitions on student performance. According to [1], student performance can be obtained by measuring the learning assessment and curriculum. However, most of the studies mentioned about graduation being the measure of students' success [2], [3].

In recent years, the situation of students in the institutions have been academically warned tended to accelerate. For example, at Can Tho University¹, in the first semester of the school year 2018-2019, the number of students who academic warned in one semester were 886 and the two semesters were 125, these number in the first semester of the academic year

2019-2020 were 986 and 196 respectively. One of the main reasons for the students' poor performance is that they have not selected appropriate courses to their competencies. These results in extension of learning term and increase of cost for their families, higher educational institutions and society as well. Therefore, predicting students' performance is an important research topic in exploiting educational data, which is of interest to many researchers [4]–[8].

Currently, there are a lot of proposed approaches to predict student performance, in there data mining is one of the most popular approaches to be widely applied in educational area. One of the most popular techniques to predict student performance is classification. There are several algorithms used for classification task such as Decision Tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machines [3]. However, the existing researches are primarily based on learning results of previous semesters to predict student performance of next semester or Current Grade Point Average (GPA), but do not analyze additional factors such as English entrance testing grades, activity incentive grades, etc. that affect their performance. Moreover, the researchers have not sufficiently compared among techniques, especially deep learning techniques with other traditionally machine learning techniques.

This study presents an approach of deep learning techniques [9] using the convolutional neural network on 1D data (CN1D) and the Long Short Term Memory (LSTM) to build a student's performance prediction model for predicting student performance in next semesters based on the course's achievement results of the previous semesters.

We analyze and introduce some techniques for data pre-processing (including Quantile Transforms and MinMax Scaler) before fetching them into well-known deep learning algorithms such as LSTM and Convolutional Neural Networks to do prediction tasks. In addition, in order to improve the predictive results, we also consider other additional factors such as entrance English testing grades, activity incentive grades etc. for the proposed model. Moreover, a comparison between deep learning techniques and traditionally machine learning ones is also conducted. Experimental results show that the proposed model provides rather accurate prediction and it can be applied in practical other cases.

In the remainder of this study, we present a literature review on studies performing on student performance prediction. We introduce the considered 17 datasets in Section III. The works

¹www.ctu.edu.vn

with the proposed techniques and algorithms are presented in Section IV. The experimental results include the model settings and performance metrics, data transformations, optimizer functions, results' comparison in Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), and factor analysis which are presented in Section V. We conduct and summarize some remarks in Section VI.

II. A LITERATURE REVIEW ON STUDENT PERFORMANCE PREDICTION

Predicting student's academic performance becomes critical demands that institutions can support students to achieve more performance in their learning. In [2], almost 70 papers were analyzed to show different modern techniques widely applied for predicting students' performance. These techniques belongs to several areas such as Artificial Intelligence, Machine Learning, Collaborative Filtering, Recommender Systems, and Artificial Neural Networks.

According to [8], predicting student performance is an important task in exploiting educational data; student's knowledge can be improved and accumulated over time. From this idea, the authors proposed an approach that uses Tensor Factorization (TF) to predict student performance. With this approach, the authors can personalize the prediction for specific student. Experimental results on two large datasets showed that incorporating prediction matrix factorization techniques is an effective and promising approach.

The authors in [10] investigated the effectiveness of transfer learning from deep neural networks for the task of students' performance prediction in higher education. Experiments were conducted based on data originating from five compulsory courses of two undergraduate programs. The experimental results demonstrate that the prognosis of students at risk of failure can be achieved with satisfactory accuracy in most cases.

The authors in [11] developed a student performance prediction system using the open source recommendation system called *MyMediaLite*. For the grade databases collected from the academic management system of a university, the authors proposed using Biased Matrix Factorization (*BMF*) technique to predict the learning results. This results can help students choose more appropriate courses.

The ability to combine the prediction techniques is also used by researchers. [12] developed a model to predict the student learning outcomes based on the combination of Taylor approximation method and Grey models to obtain the most optimal predicted values by multitudes approximate calculation to improve the predicted accuracy of two grey models. Research results can help teachers and educational managers have appropriate solutions to improve the academic results of students who have unstable learning process. In addition, [13] used Collaborative Filtering, Matrix Factorization and Restricted Boltzmann Machines techniques to systematically analyze data collected from a university. The results showed that Restricted Boltzmann Machines technique predicts students' academic results better than the remaining techniques.

In fact, collaborative filtering algorithms are commonly used in recommendation systems due to their simplicity and

effectiveness. However, the sparsity of the data limits the effectiveness of these algorithms and it is difficult to further improve the prediction results. Therefore, the models that combine collaborative filtering algorithms with deep learning techniques are more interested. [14] proposed a model based on quadratic polynomial regression model to obtain more accurate latent features by improving the traditional matrix factorization algorithm. Then, the latent features are the input data of the deep neural network model. The experiments on three datasets showed that the proposed model improves the prediction efficiency very well compared to traditional ones. Some other approaches combining collaborative filtering model with deep learning are also proposed by [15]. With this approach, during the prediction period, a feed-forward neural network is used to simulate the interaction between the user and the item, in which the feature vectors in pre-process used as input to the neural network. The experiments based on two datasets of *MovieLens* with one million samples and *MovieLens* ten million features to verify the effectiveness of this method and gave very feasible results. [16] also presented a review on machine learning based approaches for predicting student performance. Other approaches can be found in [17]–[20].

The problem of student performance prediction has been taken into account in numerous previous research using machine learning theory but factor analysis for student performance prediction based on explanation models and data transformation techniques are still the gap for improvements. This research aims to create a new approach that leverages Deep learning. Especially, the Long Short Term Memory (LSTM) can be used with time-based features. This study includes several contributions as the following:

- Deep learning models (LSTM, CNN) with a shallow architecture are leveraged to do student performance prediction tasks. As shown from the results, deep learning techniques can produce feasible prediction scores.
- Various optimizer functions are tested to choose an appropriate one for the considered regression problem.
- Data transformation techniques are also considered to enhance the performance of deep learning models. The feature values which are greater than 1 cause poor performance for deep learning model. By using and testing various data pre-processing techniques, we found that regression tasks with deep learning can converge sooner and also archive a better result.
- We investigate and consider various 17 datasets related to a vast of majors and study fields for the comparison in a multidisciplinary university. Based on the time to divide the training and testing data, we obtain the various ratio between the training set and test set to evaluate the difference in the prediction performance.
- A variety of model explanations are brought to analyze factors which can influence on student performance. From analysis results, educational managers can propose appropriate policies and strategies to support their students.

TABLE I. INFORMATION ON THE NUMBER OF SAMPLES OF EACH STAGE AND PERCENTAGE OF TRAINING DATA OF THE CONSIDERED EDUCATIONAL UNITS AND DEPARTMENTS

Dataset	#Train Instances	#Test Instances	%Training ratio	Total
Education	292,297	78,987	78.73%	371,284
Mekong Delta Development Research Institute	27,795	10,206	73.14%	38,001
Agriculture	294,694	179,042	62.21%	473,736
Biotechnology Research & Development Institute	46,556	30,075	60.75%	76,631
Physical Education	26,318	6427	80.37%	32,745
Engineering Technology	418,835	21,4710	66.11%	633,545
Information and Communications Technology	132,907	86,901	60.47%	219,808
Natural Sciences	79,368	42,121	65.33%	121,489
Rural Development	101,039	102,994	49.52%	204,033
Environment & Natural Resources	125,659	83,687	60.02%	209,346
Economics	518,392	171,538	75.14%	689,930
Foreign Languages	125,882	56,559	69.00%	182,441
Social Sciences and Humanities	96,491	47,469	67.03%	143,960
Aquaculture & Fisheries	109,637	49,498	68.90%	159,135
Law	155,099	54,194	74.11%	209,293
Political Science	33,493	30,009	52.74%	63,502
Total (full dataset)	2,584,462	1,244,417	67.50%	3,828,879

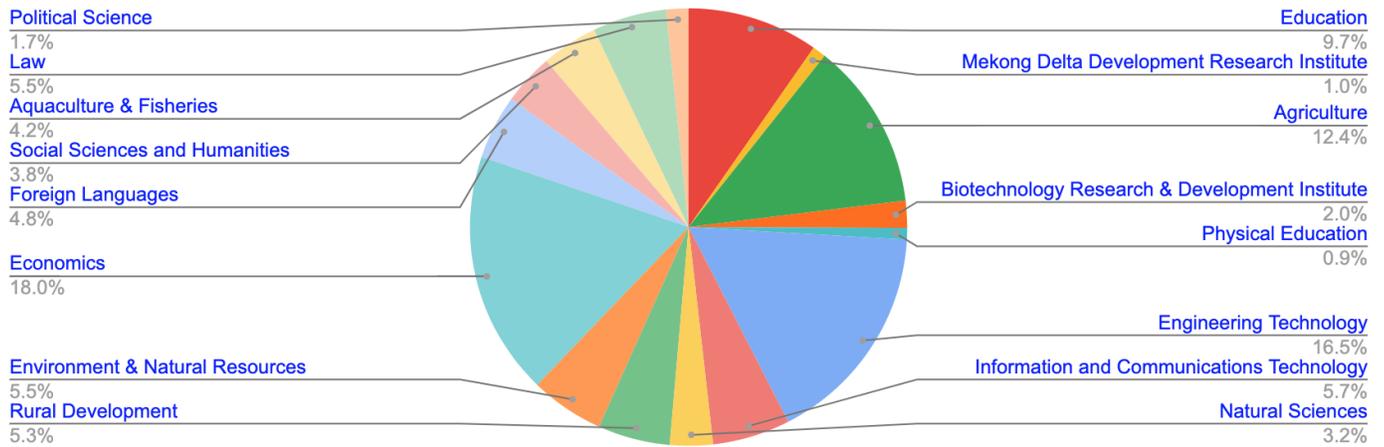


Fig. 1. Percentages of Educational Units and Departments Comparing to the Full Dataset

III. DATA DESCRIPTION

In order to evaluate the proposed model, we have collected real data at a multidisciplinary university, a case of Can Tho University, Vietnam. However, the model can be applied to other universities, schools, colleges as well. The collected data relates to students, courses, marks, and other information from the year 2007 to 2019 with 3,828,879 records, 4,699 courses (subjects), and 83,993 students. Data distributions are described in Table I with information on samples and the ratio for training of educational units/department/institutes at the considered multidisciplinary university.

The set of datasets consists of student performance from 16 academic units (faculties/colleges/schools) that belong to Can Tho University. For each unit, we separate the data into two parts, one of them for the training stage and the remaining for the test stage. Because of data division based on periods (from 2007 to 2016 for training, and from 2017 to 2019 for testing), the size of data for training and testing of each unit is different. Adding to these 16 academic units, we also evaluate our proposed method on the full dataset which includes all

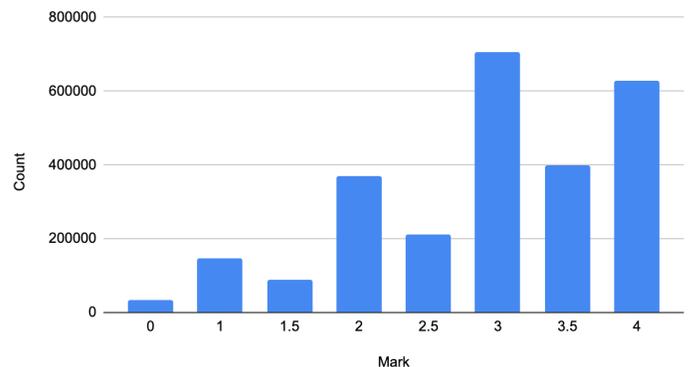


Fig. 2. Distribution of Mark Levels on Training Set of the Full Dataset.

samples from 16 academic units. Fig. 1 reveals the percentages of each units comparing to the full dataset. Economics dataset occupies the highest value with 18% and the lowest belongs to Physical Education with 32,745 samples constituting 0.9%.

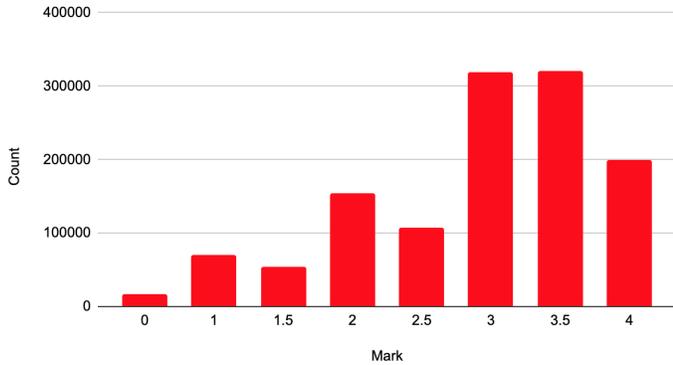


Fig. 3. Distribution of Mark Levels on Test Set of the Full Dataset.

The distribution of mark levels of the full dataset for training and testing described in Fig. 2 and Fig. 3, respectively. For these distributions, most of the marks are greater than or equal medium level as 2 (89.7% for full training dataset and 88.6% for full testing dataset). The distribution is similar to most units in the university, for instance, the mark level distribution of Engineering Technology dataset described in Fig. 4 and Fig. 5, respectively.

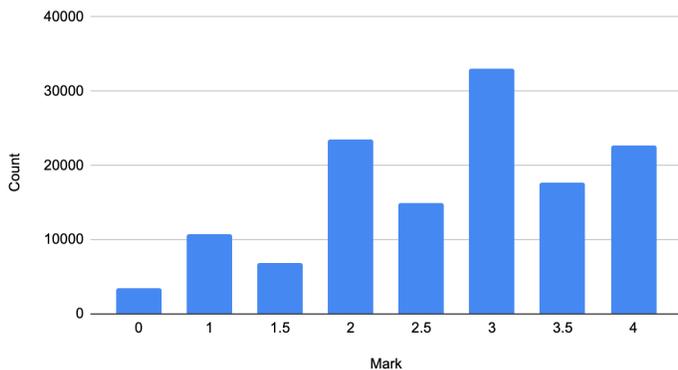


Fig. 4. Distribution of Mark Levels on Training Set of Engineering Technology Dataset.

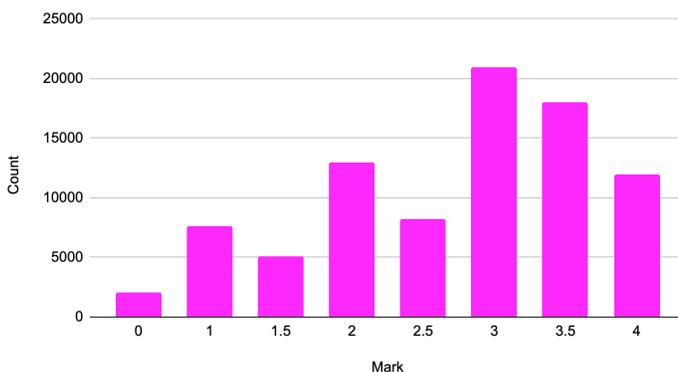


Fig. 5. Distribution of Mark Levels on Test Set of Engineering Technology Dataset.

IV. A PROPOSED APPROACH BASED ON DEEP LEARNING FOR STUDENT PERFORMANCE PREDICTION

First, we collect real datasets at the Student Management System of a university, then data is pre-processed to remove noise, redundant attributes, etc. Next, we divide the data for training and testing in term of “time order” which means that we use the “studied courses” to predict the “to be studied courses”. For example, we have used data from 2007 to 2016 for training, and data range from 2017 to 2019 for testing. The purpose of this division is to use courses results in the past (history data) to predict results of course in the future. In order to evaluate the efficiency of the prediction model, “the future” in this context is referred to data of the year from 2017.

A. Data Pre-Processing and Transformation

Since the dataset collected from the Student Management System have a lot of information, we have pre-processed them as described in the following steps:

- **Step 1:** Remove redundant attributes such as Student Name, Course Name, Lecturer Name, etc.
- **Step 2:** Remove redundant/noise records such as the courses which are registered by the student but have not been taken examination (i.e., the null marks), exemption courses, etc.
- **Step 3:** Remove the courses which have not enough registration (in some universities, if the courses are registered by less than 15 students, they will be removed).
- **Step 4:** Transform the text values to numeric values and other formats.

After carefully analyzing the data, we have selected the input attributes for learning model as described in Table II. This selection based on pre-experimental results and previous analysis in predicting student performance [11], [21].

With various data distribution of obtained different attributes, we suggest using Quantile Transformation (QTF) and MinMaxScaler (MMS) [25] for generating and convert all values to the value range where deep learning algorithms can converge.

QTF, a non-linear transformation, is considered a strong preprocessing technique because it reduces the effect of outliers. Values in new/unseen data (for example, a test/validation set) that are lower or higher than the fitted range are set to the bounds of the output distribution. As shown from Fig. 6 with an example from samples of Rural Development, before data is transformed, data range and distributions of each feature have great differences. Data transformed with QTF with all features range from 0 to 1 (Fig. 6b). Fig. 6a exhibits the result of scaler for each feature also enables its distribution to become more normal distribution.

MMS is also used for creating bins for images. This algorithm scales each feature to a given range with formulas 1 and 2:

$$X_{std} = \frac{X - X.min}{X.max - X.min} \quad (1)$$

TABLE II. INPUT ATTRIBUTES. EACH SAMPLE HAS THE LABEL WHICH IS A MAGNITUDE OF MARK/SCORE AS OUTPUT CORRESPONDING TO THAT SAMPLE. THE MAGNITUDE OF MARK/SCORE IS THE TARGET FOR THE PREDICTION TASKS.

Feature names	Description
CGPA	Average marks of the courses credited (passed) in all previous semesters
CGPA-PreSemester	Average marks of the courses credited (passed) in the preceded semester
CourseID	ID of the course
Credits_earned	The number of credits which the student earned from previous semesters
EnglishMark_11	Basic English Course- Level 1
EnglishMark_12	Basic English Course - Level 2
EnglishMark_13	Basic English Course - Level 3
EntranceMark_s1	Entrance mark for Course 1
EntranceMark_s2	Entrance mark for Course 2
EntranceMark_s3	Entrance mark for Course 3
EntranceYear	Course of the student, e.g. 2007, 2009, etc.
Faculty	Student belongs to a faculty (e.g. Economics, etc.)
FieldOfStudy	Student's Field of study
Gender	Student gender
GPA_Semester	Average marks of the courses in the preceded semesters
HighSchoolPlace	High school where the student graduated
LecturerID	Lecturer ID who taught the course
Mark_RecordedTime	The time when the mark was recorded in the system
No._Semester	Student's Semester order (eg. 1st term, 2nd term, etc.)
NumberOfCredits	The number of credit of the course
StudentID	Student ID

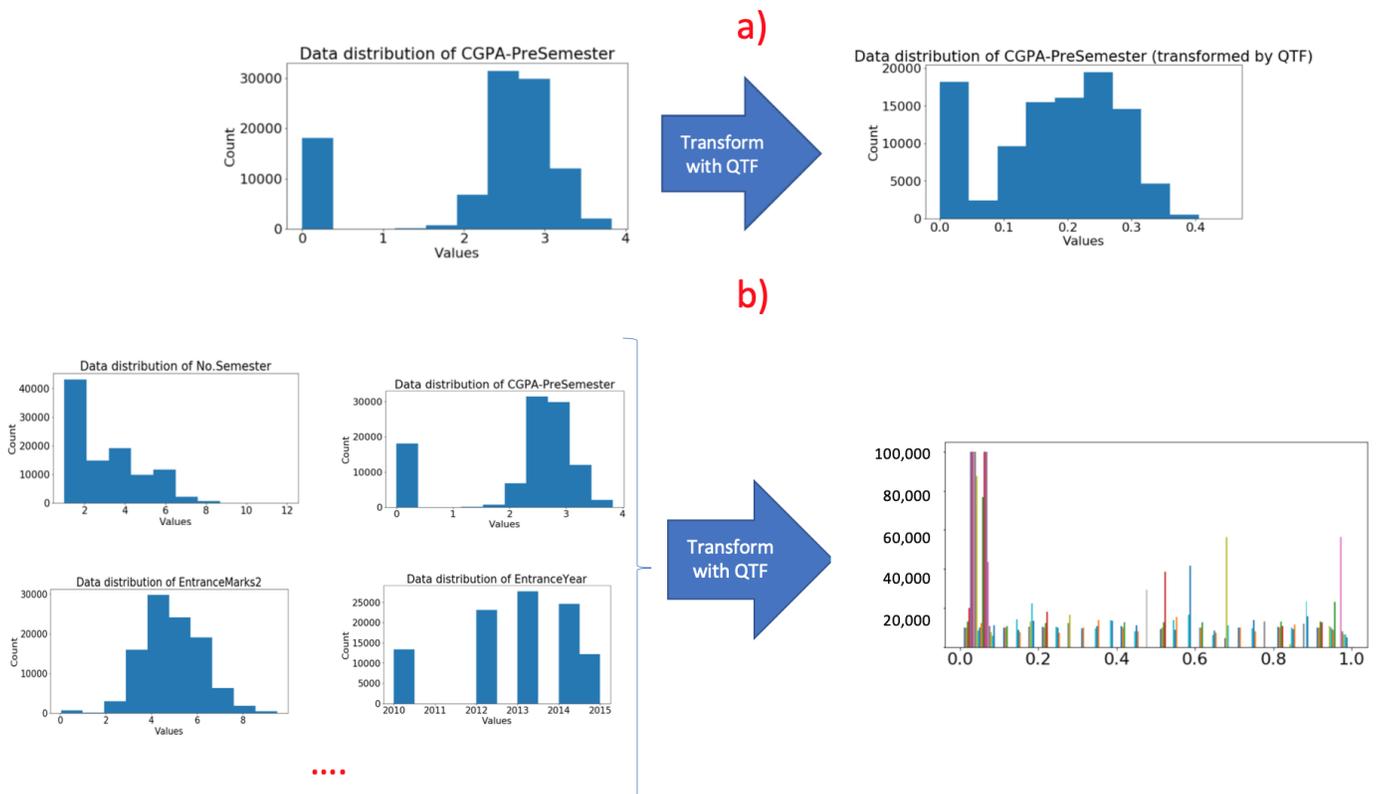


Fig. 6. Data Distributions before and after using QTF on the Training Set of Rural Development Dataset. (a) The Results of before and after Transforming a Feature (CGPA-PreSemester) as an example. (b) An Illustration of Data Distribution with Various Data Ranges from some Features before applying the Transformation and Data Distribution of all Features Transformed using QTF with the Scale from 0 to 1.

$$X_{scaled} = X_{std} * (max - min) + min \quad (2)$$

These algorithms are proven as an efficient method in classification tasks in [22]. In this study, the experiments

also reveal promising results comparing to original data with regression tasks. The scaler is learned from the training set and applying to the test set.

B. Proposed Models

Two deep learning and a robust regression (Linear Regression) algorithms are carried out to run prediction tasks.

OPERATION		DATA DIMENSIONS	WEIGHTS(N)	WEIGHTS(%)
Input	#####	21 1		
Conv1D	\ /	-----	256	17.4%
relu	#####	19 64		
Flatten		-----	0	0.0%
	#####	1216		
Dense	XXXXX	-----	1217	82.6%
sigmoid	#####	1		

Fig. 7. The Proposed CN1D Architecture

The convolutional neural network (namely, CN1D) receives 1D data with 21 features, then passing through a stack of one convolutional layer with 64 kernels of 3 (stride 1), followed by a ReLU activation function used after each convolution (shown in Fig. 7).

OPERATION		DATA DIMENSIONS	WEIGHTS(N)	WEIGHTS(%)
Input	#####	1 21		
LSTM	LLLLL	-----	22016	99.7%
tanh	#####	64		
Dense	XXXXX	-----	65	0.3%
sigmoid	#####	1		

Fig. 8. The Proposed LSTM Architecture

The Long Short Term Memory (LSTM) includes 64 Tanh units and one time step (Fig. 8). Both the CN1D and the LSTM produce output by a sigmoid function (Equation 3). The output of the sigmoid function ranges from 0 to 1, so this output then multiplied by 4 to corresponding the grades scale ranging from 0.0 to 4.0 for the mark prediction.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

V. EXPERIMENTAL RESULTS

A. Learning Settings and Performance Metrics

All networks with deep learning models deploy either Adam algorithm or Root Mean Square Propagation- RMSprop algorithm [23] as the optimization functions with a learning rate of 0.0001, a batch size of 16000 running to 500 epochs. In order to reduce overfitting, we used early stopping with the epoch patience of 5. If the loss cannot be reduced after consecutive epochs, the learning will be stopped. The scaler algorithm learns from the training set and transforms both training and test sets.

The regression performance is measured using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) averaged 5 times of run on the test set. The root mean squared error and mean absolute error (MAE) are calculated by equations (4) and (5), respectively.

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \tag{4}$$

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{5}$$

where, y_i is the true value (Grades: 0.0-4.0 scale), and \hat{y}_i is the predicted value.

The experimental results are presented as follows. The results with various scalers are illustrated in Section V-B where we show QTF as an appropriate solution for data pre-processing for regression tasks. Then, the selected scaler is run with two optimizer functions including Adam and RMSprop for the comparison in Section V-C. We run the experiments using deep learning and linear regression with the best scaler and optimizer function in Section V-D on all 16 datasets including department and institutes of the considered university and also carried out the prediction on the full dataset which merged from 16 datasets of educational units.

B. Scalers Enhance the Performance of Deep Learning

Various scalers provide the results shown in Table III. It is clear to see that the scalers are able to improve the performance of deep learning algorithms. QTF which can be the best choice among the considered scalers reveals the highest performance on 15 out of 16 datasets for CN1D and all datasets for LSTM. Fig. 9 and Fig. 10 also exhibit a clear view of the improvement. CN1D benefits from the scaler with a significant improvement.

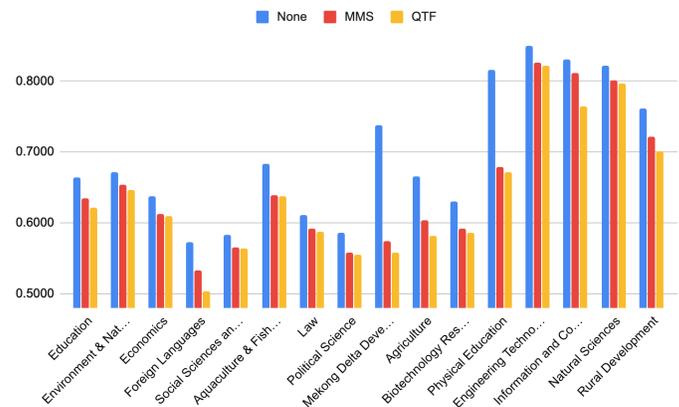


Fig. 9. MAE Comparison using Various Scalers with LSTM

C. Root Mean Square Propagation (RMS) Optimizer with Deep Learning Regression

Table IV presents the result comparison between RMSprop and Adam optimizer function. We obtain the improved performances by using RMS on 14 out of 16 considered datasets while the difference in the performance is trivial for other two datasets. The average improvement on all datasets is estimated at 3.3%. Engineering Technology dataset achieves the most improvement with reducing 0.0660 in MAE while some of the other datasets such as Environment & Natural Resources, Aquaculture & Fisheries and Information and Communications Technology can reach better results by over 0.04 in MAE.

TABLE III. DETAILS OF STUDENT PERFORMANCE PREDICTION RESULTS (IN MAE) WITH LSTM AND CN1D USING VARIOUS SCALERS. THE RESULTS IN BOLD TEXT ARE THE BEST PERFORMANCE CORRESPONDING TO EACH DATASET AND EACH CLASSIFIER

Dataset	CN1D			LSTM		
	None	MMS	QTF	None	MMS	QTF
Education	0.8595	0.6231	0.5847	0.6637	0.6341	0.6220
Environment & Natural Resources	1.0308	0.6588	0.6130	0.6716	0.6531	0.6461
Economics	1.1129	0.6122	0.6098	0.6375	0.6130	0.6092
Foreign Languages	3.2603	0.5382	0.4961	0.5736	0.5336	0.5031
Social Sciences and Humanities	3.0476	0.5711	0.5793	0.5827	0.5653	0.5647
Aquaculture & Fisheries	1.1499	0.6502	0.6471	0.6835	0.6393	0.6377
Law	1.0574	0.6017	0.5675	0.6112	0.5924	0.5878
Political Science	1.0017	0.5565	0.5547	0.5868	0.5581	0.5558
Mekong Delta Development Research Institute	3.1145	0.5803	0.5684	0.7377	0.5746	0.5587
Agriculture	0.9733	0.5995	0.5828	0.6658	0.6042	0.5814
Biotechnology Research & Development Institute	0.8133	0.6016	0.5980	0.6308	0.5928	0.5855
Physical Education	3.0487	0.6942	0.6853	0.8159	0.6795	0.6720
Engineering Technology	1.3427	0.7761	0.7487	0.8491	0.8267	0.8217
Information and Communications Technology	2.7080	0.7728	0.7285	0.8306	0.8108	0.7637
Natural Sciences	1.2244	0.8016	0.7989	0.8210	0.8008	0.7972
Rural Development	1.2134	0.7302	0.6936	0.7606	0.7220	0.7004

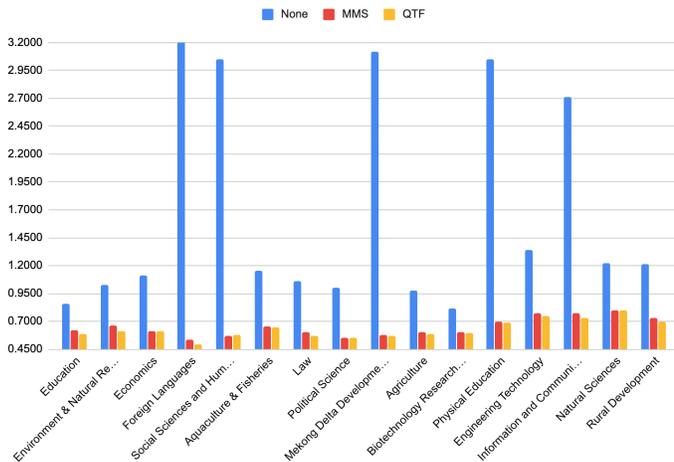


Fig. 10. MAE Comparison using Various Scalers with CN1D

D. Student Performance Prediction with Shallow Deep Learning Architectures

The performance of three learning models with the RMSprop optimizer and QTF scaler is shown in Table V. We almost achieve the best performance with deep learning (13 datasets in MAE and 12 datasets in RMSE out of 16 datasets, respectively).

CN1D holds the first place in both MAE and RMSE with achieving the best results on 9 datasets in MAE and 10 datasets in RMSE. The best MEA is obtained on Foreign Languages dataset while the worst in the prediction results are for Engineering Technology dataset. Using the CN1D model, 11 datasets have MAEs which are lower than 0.6 while there are two datasets which get high MAEs being greater than 0.7. On the other hand, the results are rather similar to the metric of RMSE. The best RMSE (0.64607) is achieved when marks come from Foreign Languages. The poor prediction results of

TABLE IV. PERFORMANCE COMPARISON (IN MAE) BETWEEN RMSPROP AND ADAM OPTIMIZER. RMSPROP GIVES BETTER PERFORMANCES WITH THE RESULTS FORMATTED BOLD. WE REPORT THE AVERAGE PERFORMANCE ON ALL DATASETS AT THE LAST ROW (AVG).

Dataset	RMS	Adam
Education	0.5882	0.6220
Environment & Natural Resources	0.5979	0.6461
Economics	0.5753	0.6092
Foreign Languages	0.4910	0.5031
Social Sciences and Humanities	0.5672	0.5647
Aquaculture & Fisheries	0.5934	0.6377
Law	0.5566	0.5878
Political Science	0.5556	0.5558
Mekong Delta Development Research Institute	0.5541	0.5587
Agriculture	0.5804	0.5814
Biotechnology Research & Development Institute	0.5763	0.5855
Physical Education	0.6837	0.6720
Engineering Technology	0.7556	0.8217
Information and Communications Technology	0.7152	0.7637
Natural Sciences	0.7740	0.7972
Rural Development	0.6989	0.7004
AVG	0.6165	0.6379

Engineering Technology samples can be explained by label distribution as shown in Fig. 4 and Fig. 5 where we can observe that the distribution of training set and test set exist some differences at the mark level of 3.5 while the low results of Rural Development dataset seem to be that the number of samples in the training set is even less than the number of samples in the test set. The models may not obtain enough data in the training set to capture the characteristics in the test set. Predicting marks from Physical Education students can be challenging because of special characteristics from this department where each student can own special talents. However, it can be that many of them usually have to focus and concentrate on various Sports competitions, and hence, he or she cannot spend more time to performs well on other subjects.

TABLE V. THE PREDICTION PERFORMANCE WITH VARIOUS LEARNING ALGORITHMS IN RMSE AND MAE USING QTF SCALER AND RMSPROP (RMS) OPTIMIZER FUNCTION. THE BEST RESULTS OF EACH DATASET ARE FORMATTED WITH BOLD TEXT

Dataset	MAE			RMSE		
	CN1D-RMS	LSTM-RMS	Linear Regression	CN1D-RMS	LSTM-RMS	Linear Regression
Education	0.57331	0.58822	0.58366	0.73217	0.74917	0.76199
Environment & Natural Resources	0.59888	0.59792	0.61096	0.75927	0.75896	0.76621
Economics	0.59218	0.57528	0.58187	0.74045	0.72979	0.73904
Foreign Languages	0.48534	0.49103	0.50141	0.64607	0.64981	0.65925
Social Sciences and Humanities	0.59204	0.56724	0.57305	0.75332	0.73727	0.73185
Aquaculture & Fisheries	0.59180	0.59338	0.59882	0.75729	0.75875	0.76099
Law	0.55460	0.55659	0.56638	0.69150	0.69367	0.70134
Political Science	0.57651	0.55560	0.56474	0.72644	0.74287	0.70574
Mekong Delta Development Research Institute	0.56779	0.55407	0.54695	0.70561	0.69634	0.69463
Agriculture	0.58055	0.58043	0.57992	0.73776	0.74044	0.74419
Biotechnology Research & Development Institute	0.53296	0.57634	0.54973	0.69808	0.77467	0.71557
Physical Education	0.67617	0.68367	0.68269	0.82931	0.83464	0.85973
Engineering Technology	0.74536	0.75564	0.75749	0.90762	0.91895	0.92585
Information and Communications Technology	0.69027	0.71518	0.73927	0.84843	0.87515	0.89998
Natural Sciences	0.67251	0.77397	0.71112	0.84147	0.99010	0.87977
Rural Development	0.71336	0.69887	0.69280	0.85659	0.84676	0.84369

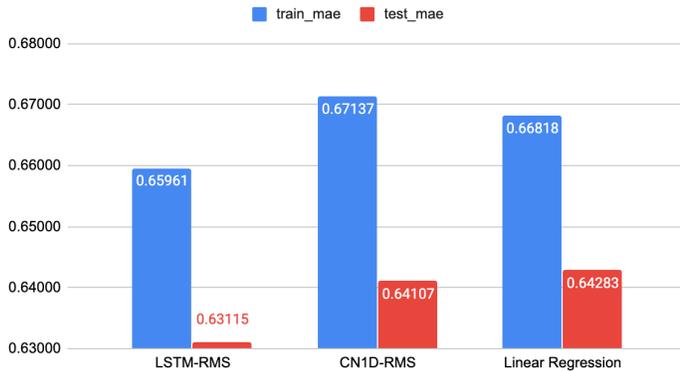


Fig. 11. The Performance in Training Phase (train_mae) and Test Phase (test_mae) of the Proposed Models on the Full Dataset in MAE

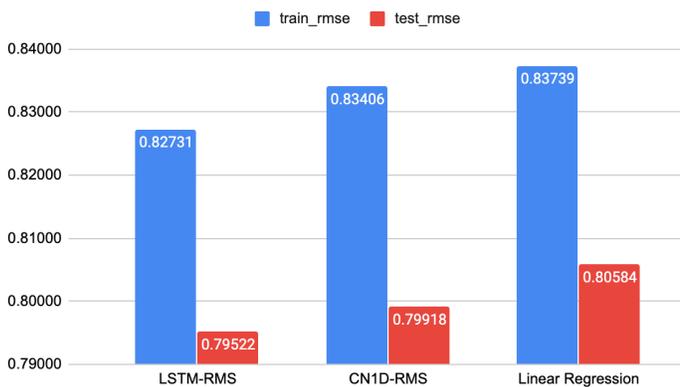


Fig. 12. The Performance in Training Phase (train_rmse) and Test Phase (test_rmse) of the Proposed Models on the Full Dataset in RMSE

With the scaler of QTF and RMSprop optimizer, we run the prediction tasks on the full dataset which contains 3,828,879 samples of all departments and institutes of the considered university. The training set includes 2,584,462 samples (67.5%)

recorded from 2007 to 2016 while the test set consists of 1,244,417 samples (32.5%) from 2017 to 2019. The results from the considered classifiers are exhibited in MAE and RMSE in Fig. 11 and Fig. 12, respectively. The LSTM holds the best in both RMSE and MAE with values of 0.79522 and 0.63115, respectively. Another deep learning algorithm also obtains better performance than Linear Regression with the results of 0.64107 and 0.79918 in MAE and RMSE, respectively. The performance in the training phase is lower than the validation phase but the difference is trivial. Comparing by MAE, the difference between training and test phases with LSTM is about 0.02846 while this value is 0.03031 for CN1D.

From these experimental results, we can observe that the proposed models could produce acceptable prediction results, thus, the system could support the students to select appropriate courses and to design suitable study plans. Moreover, student performance prediction enables lecturers as well as educational managers to indicate what students should be monitored and supported to complete their programs with the best results. These supports can reduce formal warnings and expulsions from universities due to students' poor performance.

E. Influence Factor Analysis for Student Performance Prediction

In order to know which features are important for the model to learn, we calculate the Pearson product-moment correlation which indicates the covariance of the two variables. This correlation is particularly helpful for regression problem. Next, we compute each variable's standard deviation. The correlation coefficient is indicated by dividing the covariance by the product of the two variables' standard deviations (Equation 6).

$$p_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y} \quad (6)$$

where

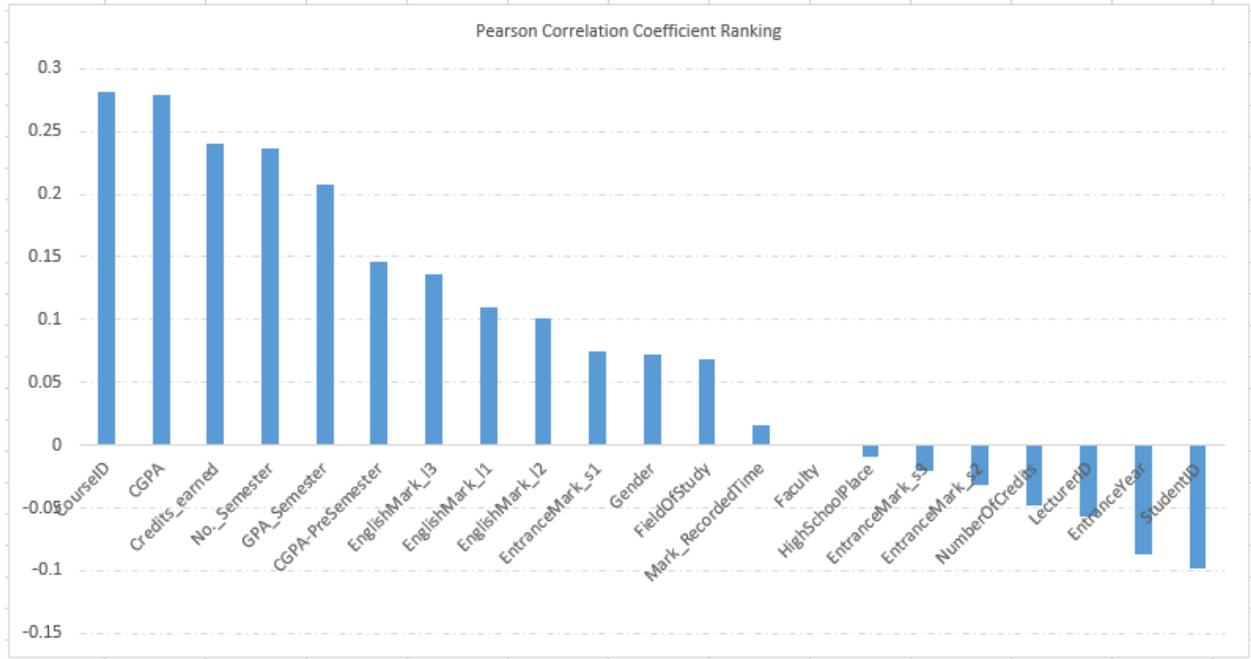


Fig. 13. Factor Analysis with Pearson Correlation Coefficient Raking. Features that their Values are Higher than Zero have Rich Information for the Model to Learn

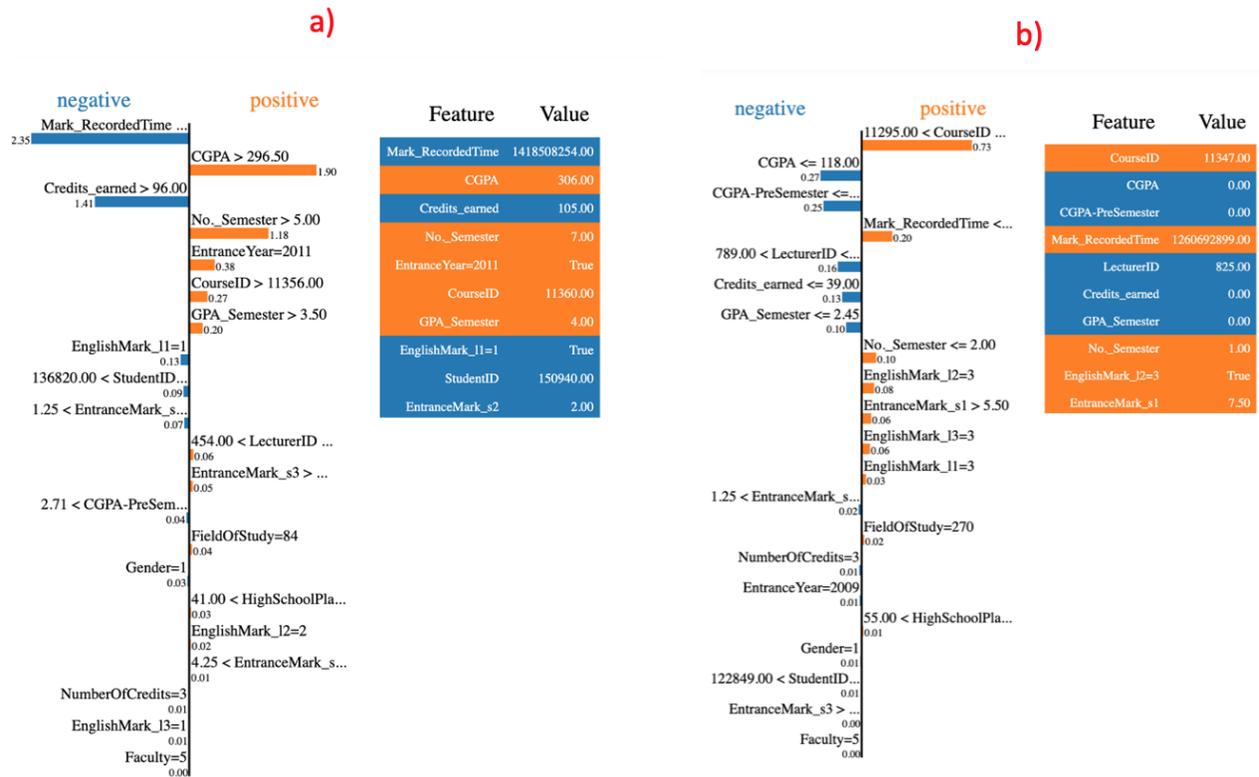


Fig. 14. Influence Factor Explanation from Local Interpretable Model-Agnostic Explanations (LIME) [24] for the Prediction Results. (a) The Prediction is 3.04 while the Real One is 3.0 (with a Low Error). (b) The Model Predicted a Mark of 3.67 while the Real Value is 2.0 (a High Error).

- p_{xy} denotes Pearson product-moment correlation coefficient.
- $Cov(x, y)$ is covariance of variable x and y.

- σ_x exhibits standard deviation of x .
- σ_y denotes standard deviation of y .

The results of factor analysis with Pearson correlation coefficient show that the CGPA and the CourseID have the most correlation to the target attribute (the mark) while the StudentID has negative effect. Other features are presented in Fig. 13.

Taking one bad prediction and one good prediction for the influence factor analysis as shown in Fig. 14, we can see that CGPA (see Table II to get details of features) contributes a positive effect on the mark to produce a good prediction (Fig. 14a) while the result in Fig. 14b considers CGPA as a negative factor so it reveals a prediction with a higher error. An observation from Fig. 14a, we noted that CGPA and the number of semesters (No_Semester) the student studied as well as the course which student was taking also contribute positive effects on the mark.

VI. CONCLUSION

In this study, we proposed deep learning models (Long Short Term Memory and Convolutional Neural Networks) to predict the student performance prediction problem in educational data mining. We analyze and propose using some techniques for data pre-processing (e.g., Quantile Transforms, MinMax Scaler) before fetching them into deep learning models and robust machine learning such as Linear Regression to do prediction tasks. Moreover, we adapt the models by using different optimizer functions including Adam and RMSprop for improving the prediction performance. Experimental results on the dataset collected from a Vietnamese multidisciplinary university's information system show that the proposed methods provide good prediction results and is expected to apply in practical cases.

Using these results, we can help both the educational managers and the students to know early warning results so that the students can have a better plan for studying. Moreover, evaluating various training courses to help the managers to propose appropriate policies.

In the future, we continue to perform experiments on other published datasets and to change the model setting for better performance as well as to compare with other approaches. Moreover, instead of using one model to predict all of the students, future studies can investigate on separated various groups of students depending on different levels of marks to create group of models for enhancing the prediction performance. Further researches should also take into account sophisticated models which can be potential to improve the performance.

REFERENCES

- [1] U. b. Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," in 2013 IEEE 5th Conference on Engineering Education (ICEED), 2013, pp. 126-130.
- [2] Rastrollo-Guerrero, Juan & Gomez-Pulido, Juan A. & Domínguez, Arturo. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences. 10. 1042. 10.3390/app10031042.
- [3] A. M. Shahiri, W. Husain, and N. a. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Computer Science*, vol. 72, pp. 414-422, 2015/01/01/ 2015.
- [4] Lorenz Kemper, Gerrit Vorhoff & Berthold U. Wigger. Predicting student dropout: A machine learning approach, *European Journal of Higher Education*, 10:1, 28-47, 2020. DOI: 10.1080/21568235.2020.1718520
- [5] Ang, Kenneth Li-Minn, Feng Lu Ge, and Kah Phooi Seng. "Big Educational Data & Analytics: Survey, Architecture and Challenges." *IEEE Access* 8 (2020): 116392-116414.
- [6] Kim, Byung-Hak, Ethan Vizitei and Varun Ganapathi. "GritNet: Student Performance Prediction with Deep Learning." *ArXiv abs/1804.07405* (2018): n. pag.
- [7] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting Students Performance in Educational Data Mining," in 2015 International Symposium on Educational Technology (ISET), 2015, pp. 125-128.
- [8] N. Thai-Nghe, T. Horvath, and L. Schmidt-Thieme. Factorization Models for Forecasting Student Performance, in Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero, C., and Stamper, J. (Eds.) *Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011)*.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015.
- [10] Tsiakmaki, Maria & Kostopoulos, Georgios & Kotsiantis, Sotiris & Ragos, Omiros. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. *Applied Sciences*. 10. 2145. 10.3390/app10062145.
- [11] T.-N. Huynh-Ly and N. Thai-Nghe, "A system for predicting students' course result using a free recommender system library of MyMediaLite (in Vietnamese)," in *Information technology conference 2013*, Can Tho University, 2013.
- [12] N. P. Hai, T.-W. Sheu, and M. Nagai, "Predicting the Student Learning Outcomes Based on the Combination of Taylor Approximation Method and Grey Models (in Vietnamese)," *VNU Journal of Science: Education Research*, vol. 31, no. 2, pp. 70-83, 2015.
- [13] Z. Iqbal, J. Qadir, A. Mian, and F. Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study," 08/17 2017.
- [14] L. Zhang, T. Luo, F. Zhang, and Y. Wu, "A Recommendation Model Based on Deep Neural Network," *IEEE Access*, vol. 6, pp. 9454-9463, 2018.
- [15] M. Fu, H. Qu, Z. Yi, L. Lu, and Y. Liu, "A Novel Deep Learning-Based Collaborative Filtering Model for Recommendation System," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1084-1096, 2019.
- [16] J. Rastrollo-Guerrero, J. A. Gomez-Pulido, and A. Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," *Applied Sciences*, vol. 10, p. 1042, 02/04 2020.
- [17] B. Sekeroglu, K. Dimililer, and K. Tuncal, *Student Performance Prediction and Classification Using Machine Learning Algorithms*. 2019, pp. 7-11.
- [18] Arto Hellas, Petri Ihanntola, Andrew Petersen, Vangel V. Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. Predicting academic performance: a systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion)*. Association for Computing Machinery, New York, NY, USA, 175-199. DOI:https://doi.org/10.1145/3293881.3295783
- [19] E. Alyahyan and D. Düşteğör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, p. 3, 2020/02/10 2020.
- [20] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Transfer Learning from Deep Neural Networks for Predicting Student Performance," *Applied Sciences*, vol. 10, p. 2145, 03/21 2020.
- [21] N. Nguyen Thai, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007, pp. T2G-7-T2G-12.

- [22] T. H. Nguyen and J. Zucker, "Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches," in 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-5.
- [23] S. Ruder, "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747, 2016.
- [24] Marco Tulio Ribeiro et al. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. <https://arxiv.org/pdf/1602.04938.pdf>. 2017.
- [25] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, 01/02 2012.