Predicting Students at Risk Using Machine Learning and Deep Learning Techniques in MOOC Online Courses

Vinit Bhosale
vinit.bhosale2@mail.dcu.ie
Department of Computing
Dublin City University
Dublin, Ireland

Prateek Sakaray
prateek.sakaray2@mail.dcu.ie
Department of Computing
Dublin City University
Dublin, Ireland

***Abstract*— Educational Data Mining(EDM) is a new growing research area since 2007, where the essence of data mining concepts are used in the educational field for interpreting and extracting useful information of the student's in the learning process. As this field of study is related to academic performance, it is necessary for universities to develop a predictive model to identify the potential risk of a student getting failed or a student getting withdrawn from the course which is also called student attrition. EDM methods are often different as it has a multilevel hierarchy and non-independence factors in educational data. In this connection, the purpose of this study is to predict and compare the result of students based on their academic performance using different models.**

Keywords—Educational Data Mining, feature selection techniques, students at risk, Virtual Learning Environment(VLE), Convolutional Neural Network (CNN), deep learning, Naive Bayes (NB), Decision Random Forests (DRF).

I. INTRODUCTION

Educational Data Mining is an area of scientific inquiry that mainly focused on developing the methods for discovering unique and effective characteristics of a student in a learning environment [1]. One of the key areas of applications of EDM involves the identification of slow learners and weak students who are likely to have low scores. Prediction of a student's academic performance is helpful in every single instructive and educational establishment for distinguishing moderate students who are probably going to come up short or have low scholarly scores [1]. As a result of which many students fail to complete their course degrees within the required periods. Also, Student attrition at university has a major impact on students, institutes, and people in society. EDM can help universities by giving a clear understanding and a pictorial representation of the specific hindrances in the student learning career phase. For example, students can fail in the advanced subjects because they face potential issues in learning the basic information from the prerequisite subjects [2].

Data mining has many techniques for dealing with a large database. Nowadays, there is a huge amount of data currently being generated thereby exceeding the human ability to analyze and extract the most useful data without the help of an automated analysis system [2][3]. From a practical point of view, EDM allows the end-users to extract knowledge from student data for improving the quality of the teaching and learning process (T&L) by developing the interventions and computational approaches, the DM techniques in the field of education can directly help to improve the quality of T&L processes. Similar approaches have been implemented successfully in business data such as e-commerce systems to improve the sales profit [2]. With an increase in the development of the internet across the globe, many emerging forms of distance learning can easily eliminate the temporal separation between two learners. As a result of which knowledge can be transmitted to all corners of the world through an online learning platform [4]. However, asynchronized distance learning is proved to have more disadvantages. For example, if the learners taking up the distance learning courses have any doubts or problems then the different ways to solve their problems involve messages or emails, it takes a longer time when compared to the face-to-face environment [4]. Therefore, learning patterns in all the educational settings are now shifting to synchronous distance learning (live) environment which has become the preferred choice for every teaching staff.

This study aims to explore the learner's learning experience with the use of educational data mining through the virtual learning environment (VLE). Such analysis gives the institutional and administrative academic community to imply new measures and strategy policies for the improvement of students who are at risk of getting failed or withdrawn from the course which will help to reduce student attrition. This study intends to investigate the effectiveness of deep learning practices for predicting students' results based on students' interaction with the resources available on virtual learning platforms.

The objective of this study was broken down into two parts:

- What are the important features for classifying students who are at risk?
- How can you predict future outcomes for students who are at risk?
- What are the different models that can be used for analysing VLE interaction of students in courses

II. RELATED WORK

Since 2011, the yearly International Conference on Educational Data Mining (EDM) and the yearly International Conference on Learning Analytics and Knowledge (LAK) have seen numerous papers submitted and displayed to grandstand the developing and quick creating field of Instructive Information Mining and Learning Analytics. The trend in the number of articles published in the 5 years from 2011 to 2015 clearly states increasing interest in this field as shown in the Fig.1 [5]

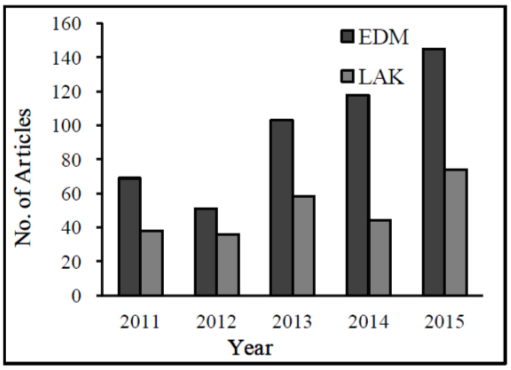

Fig.1. Articles published/submitted in EDM & LAK [5].

There are few major topics in which the researchers have outlined in the EDM 2014 conference which are listed below as shown in Fig.2:

1. Behaviour Detection.
2. Skill Estimation.
3. Game-based Learning.
4. Student Modelling.
5. Performance Prediction.
6. Q-Matrix.
7. Adaptive Learning.
8. Attrition Risk Prediction field.

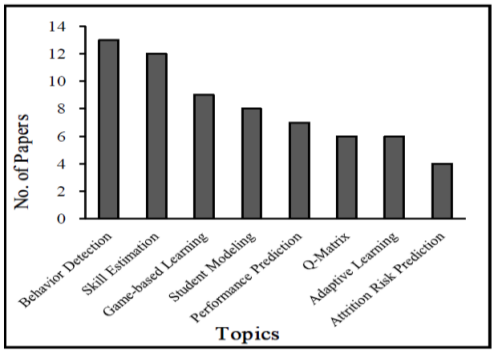

Fig. 2. Popular Topics in EDM 2014 [5]

Study in [6] explored new challenges from the EDM such as a student can learn from different sources for different topics so building a model that can analyse the outcome of the first source which can improve the quality of the second source. So in [6], the challenge would be connecting two systems from which students will be learning and change the behavior of the second system according to student's gain from the first system.

Research [7] analysed student's data related to core subjects in Computer Science (CS) to learn the importance of course content that determines the success of the student. Researchers have analysed a dataset of the historical log files from the DMIT about all courses students passed over 4 years. Dataset consists of 13640 records with 21 attributes. 1040 students completed a total of 1271 different courses completing a total of 64905 credits from which 64% of these credits are of CS category. The objective here was to better understand the students' success patterns given the core subjects to the rest of their studies. First, they have implemented a correlation analysis with Bonferroni Correction. In [7] relationship mining is used to find out the courses that have the highest correlation to the student getting success, which means if a student has performed well in certain courses then he or she can probably score well in final exams too. Second, researchers performed cluster analysis using Iterative Relocation clustering algorithm which concludes that certain courses are important for core CS. Students who performed well in these in terms of grades also perform well in terms of several courses. Third, they used Multilayer Perceptron (MLP) neural networks to predict the grades and credits of the students by providing the grades of core subjects only. Input data is provided of the students who have completed half of the core course and as output data for each student the mean grade and mean number of credits per semester are predicted.

The model built in [8] predicts the graduate level performance using the indicators of undergraduate- level performance. They have analysed records of 181 students of Computer Science from ETH Zurich, Switzerland collected over 8 years with 81 attributes. They have used a linear regression model with different variable selection techniques to examine the power of under-graduate indicators and found out the third year GPA is the most significant variable. While answering research questions they have used different data mining methodologies including cross-validation which prevents information leakage from training and testing data, bootstrapping to identify the explanatory variable. They also compared the results of different statistical estimation including AIC, BIC, and adjusted R^2 statistics.

[9] compared 2 different classification techniques Naive Bayes Classification to predict the performance based on students' academic data and Neural Networks by considering factors other than academic data. In conclusion,

it confirmed that past performances students have a significant influence on a student's academic performance.

In [12] performance of students is analysed on the same virtual learning platform used in this research where the LSTM model is used to predict the pass/fail result. This was applied weekly. They started with the 5 weeks students' interaction data and continue to analyze until 38 weeks. This research helped to analyze study patterns and also to predict early students dropout.

III. METHODOLOGY

A. *DATASET UNDERSTANDING*

This part of the research represents the exploratory analysis of practical teaching and assessment of the course modules in the Virtual Learning Environment(VLE). We have used the Open University Learning Analytics Dataset (OULAD) which was collected from Kaggle. It contains the data about courses, students, and their interactions with the Virtual Learning Environment(VLE) [10]. There are seven selected courses called modules for which the presentations are held at the end of each course module in February and October respectively as shown in Fig. The dataset consists of different student features, attributes, and tables connected using unique identifiers [10].

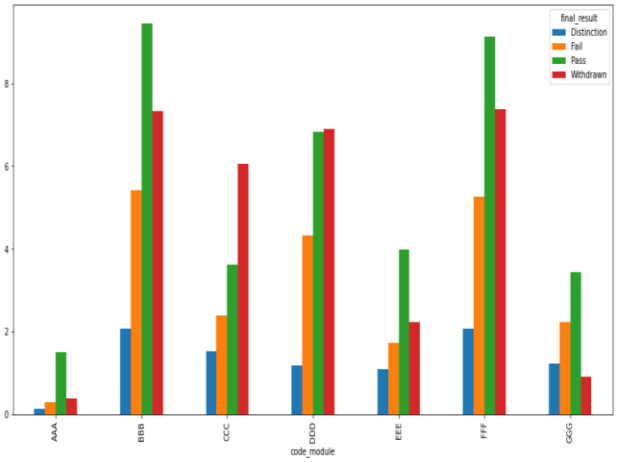

Fig. 3. Distribution of student results for each module.

There are major six instances in this dataset as listed in Table 1 with their respective entries which are recorded against the students who had appeared or enrolled for the courses and those who gave the presentations for the respective modules. The total number of attributes are 43 which some of which included features like (code_module, code_presentations, length of the module presentations in days, assessment_id, assessment_date, gender, region, highest_education, study_credits, disability, final_result, score). There were few missing attributes typically encountered in the fields of score and exam modules. This dataset included all the details of the student who took the synchronous learning behavior through VLE which recorded all the learning activities of the student with the online material which he/she accessed for studying for the exams. Features like sum_clicks on the materials were recorded since the start of the course module and were presented with a unique identification number for each student.

Instances	**Entries**
Student in courses	32953
Course- presentations	22
VLE Pages	6364
Registration entries	32953
Assessments	206
Assessment entries	17912

Table 1. Number of instances in the dataset.

The score for assessments was calculated and marked in the range of 0 to 100. The score lower than 40 is interpreted as a failure. The below Fig.4 gives the overall distribution of the students who got Distinction, Pass, Fail, and Withdrawn as their grades in the exams.

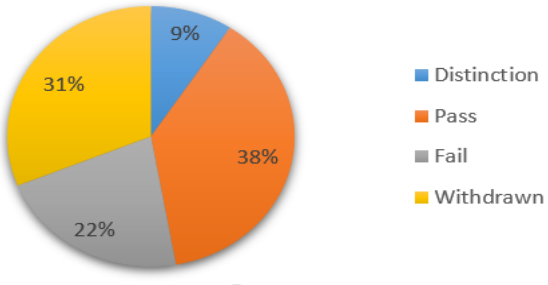

Fig. 4. Distribution of student results

In this study, we have selected results as our target variables for classifying students who are potentially at risk of getting failed after their final presentations.

B. *MACHINE LEARNING MODEL*

In this research we have 4 classifiers in which the result of the student is present which are Distinction, Pass, Fail and Withdrawn. As there are 4 classifiers this is treated as a multi-class classification problem. So, the target variable for this research is the final result of the student which can be either classified in any of these four categories. We aim to use Random Forests(RF), Decision Tree (DT), Naïve Bayes (NB), and Logistic Regression (LR) and Convolutional Neural Network (CNN) as our baseline models for evaluating results for multiclass classification problem.

1) Data and Experimentation

Table 2. presents a summary of the given dataset reviewed during the selection step. According to [13] the data can be divided into three parts: demographic data, assessment, and VLE interaction over a course duration of 9 months from 2013 to 2014. Data includes the results of the assessments and scores submitted by the students during their course curriculum. VLE clickstream data of the students from the major area for the researchers to build and evaluate various models for predicting students' performance and student dropout results. Table 2. gives the overall summary of various features available in the OULAD dataset. The dataset comprises 7 different CSV files that require significant pre-processing and transformation to extract features before building the model. We aim to select the features listed in Table 2. that are used for building models for predicting potential students who are at risk of failing. The aggregated average clicks per student were processed in two different years 2013B and 2014B to visualize the students' weekly interactions with course content and studies.

Attributes Category	**Number Attributes**	**Type**
Student Registration Information	3	Categorical
Student Demographic Information	5	Categorical
	16	Numeric
Assessments Information	3	Numeric
Number of weeks	37 (2013) 38 (2014)	Numeric
Number of visits for each VLE activity_type	16 (2013) 17(2014)	Float
Cumulative clicks for each resource type (dummies)	16 (2013) 17(2014)	Float

Table 2. Summary of dataset after feature engineering.

2) Pre-processing and Transformation

The dataset was procured in a raw structured format consisting of different data files. The VLE log-file data were analysed and computed to generate new features catering to the various activities that students performed with the VLE. These computed records were first split into two major distinct sets for both 2013B and 2014B year of course. All the features for both these years were formulated by processing the provided data tables in the dataset. Features in the raw dataset were recorded on the day-wise distribution of the students' interaction with VLE which was then computed in a week-wise manner with each week detailing the same activity features. Each week consisted of a homogeneous set of students, that is, a student in week i was also present and enrolled for a course in week (i-1) and so on. Each students' information was gathered and identified based on the unique student ID in the data. Similarly, we ignored the students who were registered for multiple courses in the two different years therefore we did not analyze the data on a course granular level.

The study intended to analyze 'pass', 'fail', 'withdrawn' instances where the 'pass' instances were mapped with the 'distinction' instances to generate one single class. The data from three CSV files namely student_registration, student_assesment, assessment was merged to generate one demographic data having instances of students registered for the year 2013B. VLE data with week wise interaction of student activities were recorded with each student's information related to his/her click patterns on course content known as resource type. New columns were formed using each resource type activity and the data related to each student in the year 2013B was populated with an aggregated number of clicks on each activity. Similarly, data for year 2014B was formulated and analysis of student interactions was mapped for both these years taking into account the split for each student by the course and year.

Weekly data consisted of instances where students had no interactions with the resource type activity for which pre-processing was done accordingly thereby mapping it to '0'. Then for each '0' value and each student, the cumulative score was calculated taking into consideration that the performance of the student would be then modeled based on his number of '0' click patterns as well.

3) Data Imbalance

Class imbalance is one of the new problems that emerged when machine learning techniques came into use for major classification and regression problems[21]. Class imbalance occurs when there are many instances of a particular class in the overall data distribution[21]. In such scenarios, standard classifiers are overwhelmed by the large distributions of unwanted classes rather than the distribution of important classes [20] [21].

Table 3. shows the distributions of the OULAD dataset with classes that are put up against each presentation year. We observed that in both the years' data was imbalanced as the number of fail and withdrawn instances was almost around 14.20% and 17.30%. Since our study majorly focused on evaluating performance for students at risk, sampling techniques like SMOTE was used which populates minority class instances by randomly interpolating pairs of closet neighbors in the minor class [21]. The level of imbalance was reduced after which models were recomputed on the new transformed set of training and testing data which achieved improvement in the results concerning target variable classification and risk prediction for student dropout/fail.

Year	Class	Counts	Percentage
2013B	Pass	18621	75.40%
	Fail	4266	17.30%
	Withdrawn	1825	7.40%
2014B	Pass	18914	77.60%
	Fail	3473	14.20%
	Withdrawn	1997	8.20%

Table 3. Distribution of data imbalance.

4) Data Mining and Evaluation

This research work aims to investigate the overall performance of the student with respect to his/her academic activities through VLE course interaction for the year 2013B and 2014B. We also aim to leverage deep learning models and machine learning models to early predict the students who are at risk. The main research question was broken down into two parts. First, what are the important features for classifying students who are at risk? Second, can you predict future outcomes for students who are at risk?

Below are the following machine learning classification algorithms which were considered for predicting the students who are at risk of failing and dropping out of the course.

a) Decision Random Forests (DRF)

Random Forests are known to be a type of ensemble method that makes predictions by computing the average predictions of several base models [13]. In this algorithm, the trees are trained independently to make individual predictions, and then results are aggregated to generate the final prediction. Three main fields need to be taken into consideration when constructing a random tree. These are (1) method of splitting the leaves, (2) type of predictor to be used in each leaf, (3) method of defining randomness into the trees [13]. Each of the trees that are generated in the forests is a weak learner and that adapts to the changes as it gets built on different parts of the dataset provided[14].

In our research, we had analysed the student dropout and risk prediction to be a multiclass classification problem. Data with each unique student ID in our dataset was analysed and classified into 4 categories namely 'Pass', 'Fail', 'Distinction', and 'Withdrawn'. We modeled each student result to be put up into weeks where his cumulative score was computed. This score was then given as an input to the algorithm for generating week wise predictions. Since Random forest is considered to be one of the most powerful classification algorithms for multilabel and binary classification, we aimed at exploring this technique on our precomputed student MOOC dataset.

b) Logistic Regression(LR)

The underlying concept of logistic regression is the logit, the natural logarithm of an odds ratio [15]. Consider an instance where the distribution of a dichotomous variable is mapped with the target dichotomous predictor variable where a test of chi-square can be used and applied [16]. We can also say that odds of both dichotomous variables are interrelated as a result of which both odds ratios can be compared and can be used to recommend or model a suggestion [15]. Generally, logistic regression is suitable for defining hypotheses about the kind of feature correlation in our dataset. Relationships between outcome variables and predictor variables can be analysed by using LR [15] [16].

The logistic regression model was applied to categorical data for both demographic and VLE data. For this, feature selection techniques were implemented and then the pre-computed data of students were used to create a split on dependent and independent variables which consisted of 6 categorical features in both the year 2013B and 2014B respectively. Our researched study focused on calculating the frequency for each of our response variables according to each enrolled student for the same course in the same year. As mentioned above we observed that a small number of fail and withdrawn percentages of students in both 2013B and 2014B year. Due to skewness in the response variable, the SMOTE technique was used to remove the imbalance in the data. 51 independent variables (X) with dummies were created and used for modeling. The dependent variable (Y) was selected to be the final results of the students for each course and each year respectively.

c) Naive Bayes (NB)

Naive Bayes classifier algorithm is majorly used for classification problems in data mining and machine learning [19]. A Naive Bayes is a simple probabilistic method that predicts class probabilities. One of the reasons why it is simple is because it only scans one time for training data required for probability generation. It also handles the missing values present in the dataset by simply omitting the correlation probabilities for the attributes when calculating the likelihood for each class[19]. The effect of the stated class is independent of other variables in the dataset.

Given a training dataset , D = {X1,X2,...,Xn} in which each X variable represents a set of values represented as Xi = {x1,x2,...,xn}. The attribute values can be discrete or continuous[19]. Each training and testing instance in our dataset for 2013B and 2014B were used to calculate the posterior probabilities after which the highest posterior probabilities conditioned on X were predicted. We used sci-kit learn library for implementing the Naive Bayes algorithm.

5) Implementation

Fig. 5. represents the architecture from an implementation perspective. The original dataset was downloaded from Kaggle as one compressed zip file.7 CSV file consists of the raw anonymized data over two years 2013 and 2014 respectively. Firstly, data pre-processing was done using python pandas and NumPy. Missing values, duplicate entries, '?' were handled. Data from 7 different CSV files were split into two categories (1) demographic data which consisted of students background information related to his academics, scores, geographical location, etc. (2) VLE data which was mainly information about each students' click history for the particular course in which student was enrolled. In this research, we aimed at analysing the click patterns based weekly and day wise as well. Machine learning classification models used week wise feed to the algorithm whereas deep learning CNN model was computed with VLE clickstream data spread across the day-wise history of click patterns of each unique student identified by unique student ID.

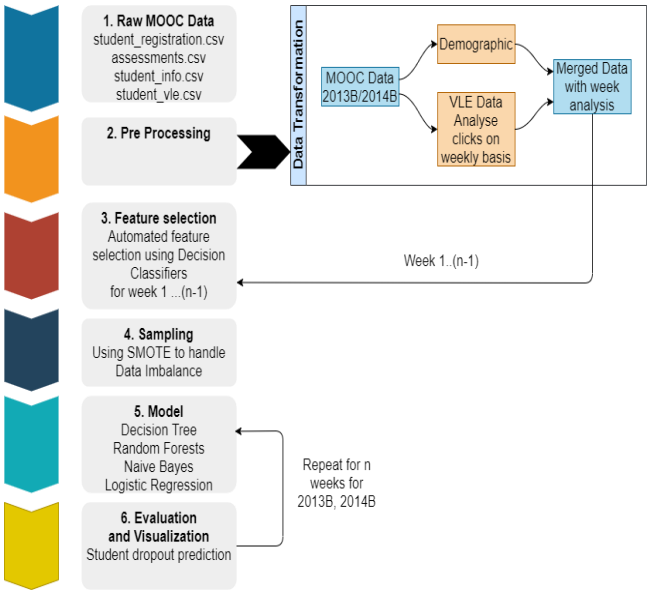

Fig. 5. Implementation architecture for the Machine Learning model.

Secondly, data was merged after pre-processing and cumulative scores for each student click activity were calculated and stored in dummy variables thereby generating new features for computation. Similarly, this data was then processed for each week in both the years 2013B and 2014B respectively. Third, the transformed dataset was imported for further post-processing. The automated feature selection technique was implemented to calculate the useful features from pre-computed data using decision classifiers which gave the feature importance matrix for using the selected number of features for modeling. Sampling technique SMOTE was used to handle an imbalance in the data which was checked against each class of response variable 'final_result'.

Finally, the performance was extracted and processed for visualization and presentation. Results were calculated based on analysis for year 2013B which was used for training purposes and testing was done against the data analysis for year 2014B respectively.

C. Convolutional Neural Network(CNN)

Our aim of this research in deep learning on the OULAD dataset is to get machines to recognize the object in the same way as our brain identifies the objects when we see a picture[20]. CNN is a part of deep neural networks which extracts the low-level features such as edges and curves to find out the patterns from the images[20]. According to the study carried out in [20], CNN can be used in face recognition, natural language processing (NLP), image and video processing, and image classification. For computers, image is just an array of numbers each object has its pattern that will help the computer to classify the object. In this study, the time series clickstream data of students on the resources are aggregated daily and graphs have been plotted to find out the study patterns of students to get a specific result.

1) CNN Layers

CNN model consists of 4 important layers which are as follows:

a) Convolutional Layer

This layer extracts features from the input images. Convolution helps to find sharpening and edge detection and other characteristics of images. This layer includes several filters whose parameters need to be learned. These filters are convoluted to input volume to compute the activation map made of neurons [17]. The output of this layer is computed by stacking the activation maps of all filters along the depth dimensions. Mathematically a convolution of two functions a and b is defined as below

$$a * b(i) = \sum_{j=1}^{m} b(j).a(i-j+m/2) \qquad (1)$$

It is a dot product of the input function and kernel function. In this research, Conv2D layer is used for modeling [17].

b) Pooling Layer

The pooling layer is used to reduce the dimensionality of data by combining the outputs of neurons

at one layer into a single neuron in the next layer without losing important features or patterns[17][18]. In this research, MaxPooling2D layer is used which uses the maximum value from each of the neurons clusters from the previous layer.

c) Flatten Layer

In this layer, a 2-dimensional matrix of features is converted in a single vector which is forwarded to a fully connected classifier[17].

d) Dense Layer

This layer is also called a fully-connected layer that connects every neuron in one layer. Results of convolution are fed to these layers which classify the image.

2) Data Generation for CNN

This section defines the steps that have been carried out to generate the clickstream data graphs for each student having a unique ID. Before plotting the images pre-processing steps implemented. First, data was merged from the 7 different CSV files then-new column of 'sum_clicks' of students were added in which clicks were aggregated on 'id_student', 'date', 'num_of_prev_attempts', 'code_module', ' code_presentation' columns. Second, categorical data is encoded into numerical data by using LabelEncoder from sklearn.preprocessing. Graphs were created in which the number of days was on the X-axis and sum of clicks was plotted on Y-axis. Labels were assigned to the graphs in the form of file names where the result of each student is given including their 'id_student', 'code_module', and 'num_of_prev_attempts'. Later the result of the student is extracted from the label and used to feed the target variable for CNN.

Now the important step was to create separate graphs for each student for each module examination. In the dataset there were many instances where a student is registered for more than one module or student has given many attempts for the same module so the interaction of the student for each examination is different. Therefore, the result of the student is also different for each unique student ID. For example, Fig. 6,7,8 represents 3 different graphs of a single unique student where it represents the results as pass, fail, withdrawn respectively. Similarly, we observed that the same student was also registered in more than one module and has given more than one attempt for final examinations. Data split has been done across training, validation, and testing with 75%, 15%, and 10% respectively.

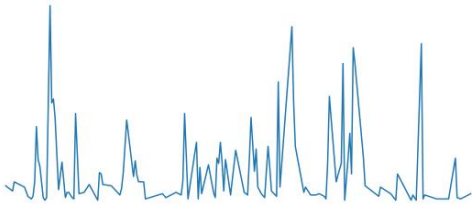

Fig. 6. Graph for passed module

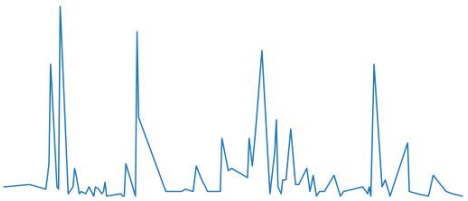

Fig. 7. Graph for failed module

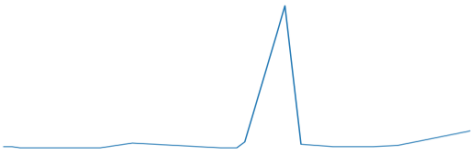

Fig. 8. Graph for withdrawn from module

3) CNN Architecture

Images generated at the previous step are of size 400*100 (W*H) are converted into grayscale images before feeding to CNN. After that target variables were extracted from the image labels. As it is a multi-class classification, target variables were binarized. Images are then converted into NumPy arrays which are later fed to the CNN model for the predictions. Fig 9. describes the complete CNN implementation architecture with all layers. The CNN architecture comprises of overall 10 layers which are shown in Fig. 9.

First, conv2D layer is added with input shape (100,400,1), hyperparameter 'kernel_size' was set to (3*3)

which was selected using the brute force approach. Parameter padding was initialized as 'valid' with 'pool_size' as (2*2).

Second, the normalization layer of 'batchnormalization' was added to the model followed by the 'Rectified Linear Unit' (ReLU) activation function. Third, the flatten layer is applied to convert a 2-dimensional matrix into a dimensional matrix. Finally, 2 dense layers with a dropout of 0.5 were added before the final dense layer where 'softmax' function is used to predict the outcomes in 4 classes. The model was compiled using optimizer as 'adam' and loss as 'sparse_categorical_crossentropy' Model is trained on 50 epochs with 'batch_size' of 32.

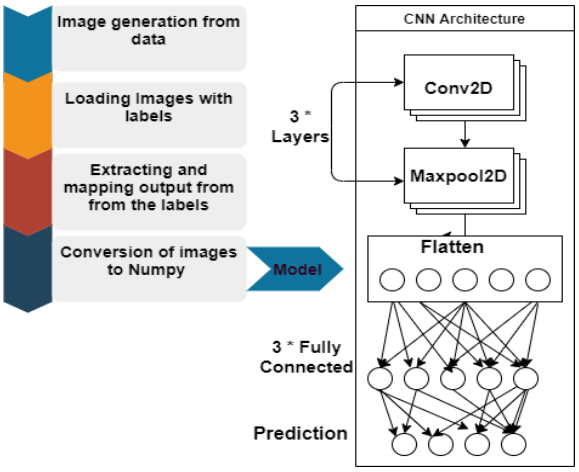

Fig. 9. CNN implementation architecture.

III. Evaluation And Results

This paper investigates the performance of various machine learning algorithms for predicting students at risk of failing and students dropping out of the course. Two different experiments were conducted to analyse the prediction results of machine learning model and the CNN model built on different categories of predictors like demographic data and VLE interaction-based. The id_student, code_module, module_presentations, and scores for each student were excluded and were not used as predictors.

Year	Model	Class	precision	recall	f1-score	support	accuracy	AUROC
2013	LR	0	0.56	0.06	0.11	3708	0.67	0.65
		1	0.5	0.13	0.2	46970		
		2	0.68	0.97	0.8	190471		
		3	0.33	0	0	1386		
	DT	0	0.89	0.94	0.92	506810	0.87	0.86
		1	0.92	0.82	0.87	506727		
		2	0.94	0.79	0.86	506546		
		3	0.8	0.97	0.88	506845		
	RF	0	0.89	0.94	0.92	126616	**0.88**	**0.87**
		1	0.91	0.82	0.86	126712		
		2	0.94	0.78	0.85	126701		
		3	0.8	0.97	0.87	126703		

Table 4. Performance metrics of machine learning model LR, DT, RF with AUC score for year 2013.

The performance metrics of results for the students in Fig depicts that the models created on demographics and VLE interactions combined achieved higher scores for RF with 0.88 accuracies and 0.87 AUC for the year 2013 as shown in Table 3. These models were used for multiclass classification where the class '0' was mapped to Pass, '1' to fail, '2' Distinction, '4' Withdrawn. Similarly, RF with 0.82 accuracies and 0.95 AUC was achieved for the year 2014 which is shown in Table 5. The other models DT and LR had shown stable performance overtraining and testing split.

Year	Model	Class	precision	recall	f1-score	support	accuracy	AUROC
2014	LR	0	0.63	0.26	0.37	61936	0.65	0.8
		1	0.45	0.1	0.16	38423		
		2	0.66	0.94	0.78	188453		
		3	0.45	0.07	0.13	15195		
	DT	0	0.65	0.8	0.71	502364	0.63	0.94
		1	0.63	0.57	0.59	501880		
		2	0.68	0.41	0.51	501807		
		3	0.62	0.78	0.69	501973		
	RF	0	0.83	0.9	0.86	125075	**0.82**	**0.95**
		1	0.84	0.76	0.8	125724		
		2	0.9	0.69	0.78	125509		
		3	0.75	0.93	0.83	125698		

Table 5. Performance metrics of machine learning model LR, DT, RF with AUC score for year 2014.

Results show that when data is analysed separately for each year yields better results as the number of entries each year corresponds to the students who enrolled for the same year courses which eliminate redundant data. Models for the year 2014 as shown in Table 5. depict that there was a slight drop in the scores as compared to the year 2013 since the data was biased as 'fail' instances recorded in the year 2014 were less as compared to 2013. Hence, each model was generated by oversampling the data thereby adding few new instances to the fail category to resolve the problem of data imbalance. After sampling the results for the year 2014 achieved are shown in Table 5. which even though is less as compared to 2013 but had shown a significant increase as

compared to the model built before using a sampling technique. Table 3. shows the distribution where we can see that there were minimum instances of the 'fail' category for predicting the results of students who are at risk and students who are on the verge of dropping out from the course.

The results obtained from CNN were not as promising as expected although different optimization techniques were applied such as 'adam', 'rmsprop', 'SGD'. Also tried with different loss functions for multi-class classification whereas 'sparse_categorical_crossentropy' was the best fit. Data augmentation is done to avoid model overfitting on training data. ReLU activation function is used for model whereas the normalization layer of 'batch normalization' is tried before and after activation function to improve the accuracy of the model. The brute force approach was used to select the hyperparameters. Still, the accuracy obtained on the training dataset is 0.55, and the accuracy obtained on the test is 0.23.

Comparing the results to previous studies that were carried out on OULAD dataset, our results are better and stable on both testing and training sets. However, both previous research work and current work have been modeled on using the best temporal features by evaluating each students' performance on various days or intervals.

IV. CONCLUSION

MOOC has now become the emerging discipline of study pattern for all the students across the globe which is giving rise to analyse student failure and student dropout rates. Many research works have mainly focused on dropout prediction. This research not only investigates the overall dropout rate but also takes into consideration all the factors of the student that can lead to his/her failure in the examinations. Various models are built on factors like demographics of the student (scores, course enrolled, geographical data of the student) and VLE clickstream data.

The research presented critical concern of students failing and students dropping out by computing different methods on weekly and daily interaction of students with the course resources. Our study compared the effectiveness of machine learning and deep learning model where the results showed that ML algorithms yielded better student performance outcomes than deep learning CNN model. The deep learning CNN model tended to capture the patterns on daily student interactions whereas the machine learning models were fed with the data which was pre-computed on weekly interaction of the students with resources to devise early intervention strategies to improve student performance. As a result, machine learning methods outperformed the CNN deep learning model. Such analysis can facilitate institutions, universities to determine the strategies to improve the performance of students.

V. ACKNOWLEDGMENT

We would also like to show our gratitude to Dr. Andrew McCarren, Dublin City University for mentoring and sharing his pearls of wisdom and so-called insights with us during this research. We are also immensely grateful for his comments on our approach for this practicum project which enabled us to gain a good understanding of the domain that we worked in together.

VI. FUTURE WORK

This study with respect to the machine learning approach does not cater to the variation in the performance of students who have repeated the course in both the 2013 and 2014 code presentation year. Therefore, analysing student behavior for his multiple course results might improve the overall learning efficiency of RF, DT, and LR. Scores can be used with the final result to map students' output to the decision target variable which can have added advantage to the given classification problem.

As the results obtained from CNN are not up to the mark images can be regenerated by doing changes in the image size. More than one image for each student can be generated to find out patterns in which 300 days of interactions will get divided. Also, the model can be trained on much deeper networks such as VGG16, AlexNet, and ResNets.

REFERENCES

[1] M. R. a. R. Bhaskaran, "A Study on Feature Selection Techniques in Educational Data Mining," *Journal Of Computing,* vol. 1, no. 1, December 2009.

[2] A. Algarni, "Data Mining in Education," *International Journal of Advanced Computer Science and Applications,* vol. 7, June 2016.

[3] Techopedia, August 2017. [Online]. Available: https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd.

[4] H.-C. Hung and I.-F. Liu, "Applying Educational Data Mining to Explore Students' Learning Patterns in the Flipped Learning Approach for Coding Education," 2020.

[5] K. Sin And L. Muthu, “Application Of Big Data In Education Data Mining And Learning Analytics – A Literature Review,” *Ictact Journal On Soft Computing*, vol. 05, No. 04, July 2015.

[6] R. S. Baker, “Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes,” *Journal of Educational Data Mining*, vol. 11, 2019.

[7] M. Saarela, “Analysing Student Performance using Sparse Data of Core Bachelor Courses,” *Journal of Educational Data Mining*, vol. 7, 2015.

[8] H. R. Heinimann, “A Model-Based Approach to Predicting Graduate-Level Performance Using Indicators of Undergraduate-Level Performance,” *Journal of Educational Data Mining*, vol. 7, 2015.

[9] H. Agrawal, “Students Performance Prediction using Machine Learning,” *International Journal of Engineering Research and Technology*, 2015.

[10] Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).

[11] Berens, J., Schneider, K., Görtz, S., Oster, S. and Burghoff, J., 2018. “Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods,” *Journal of Educational Data Mining, Volume 11, No 3, 2019*.

[12] N. Aljohani, A. Fayoumi and S. Hassan, "Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment", Sustainability, vol. 11, no. 24, p. 7238, 2019. Available: 10.3390/su11247238.

[13] Denil, M., Matheson, D. and De Freitas, N., 2014, January. Narrowing the gap: Random forests in theory and in practice. In International conference on machine learning (pp. 665-673).

[14] Jha, N.I., Ghergulescu, I. and Moldovan, A.N., 2019. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. In CSEDU (2) (pp. 154-164).

[15] Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14.

[16] El-Habil, A.M., 2012. An application on multinomial logistic regression model. Pakistan journal of statistics and operation research, pp.271-291.

[17] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review", *Neural Computation*, vol. 29, no. 9, pp. 2352-2449, 2017. Available: 10.1162/neco_a_00990 [Accessed 16 August 2020].

[18] "Convolutional neural network", *En.wikipedia.org*, 2020. [Online].Available: https://en.wikipedia.org/wiki/Convolutional_neural_network. [Accessed: 16- Aug- 2020].

[19] Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M.A. and Strachan, R., 2014. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert systems with applications, 41(4), pp.1937-1946.

[20] C. Chatterjee, "Basics of the Classic CNN", Medium, 2020. [Online]. Available: https://towardsdatascience.com/basics-of-the-classic-cnn-a3dce1225add. [Accessed: 17- Aug- 2020].

[21] Liu, T.Y., 2009, August. Easyensemble and feature selection for imbalance data sets. In 2009 international joint conference on bioinformatics, systems biology and intelligent computing (pp. 517-520). IEEE.