Int. J. Nonlinear Anal. Appl. 13 (2022) 1, 1329-1339
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.5710

Multi-class LDA classifier and CNN feature extraction for student performance analysis during Covid-19 pandemic

Rasheed Mansoor Ali S[a,*], S. Perumal[a]

[a]Department of computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India

(Communicated by Madjid Eshaghi Gordji)

Abstract

In our modern world, education is essential for developing high moral values and excellence in individuals. But the spread of Covid-19 widely affects the student's education, the majority of students have continued their education via online learning platforms. The academic performance of students has been sluggish across the globe during this pandemic. This problem is solved using a multiclass Linear Discriminant Analysis (LDA) and Convolutional Neural Network (CNN) model which predicts the student learning rate and behavior. This research aims to classify the students' performance into low, medium, and high grades in order to assist tutors in predicting the low-ranking students. The student data log is collected from the Kaggle student performance analysis dataset and pre-processed to remove the noise and non-redundance data. By analyzing the pre-processed data, the CNN extracts feature that are based on student interest and subjective pattern sequences. Then extracted features are filtered by the Minimum Redundancy Maximum Relevance(mRMR) method. mRMR selects the best features and dilutes the least one which handles each feature separately. The feature weights are measured by Stochastic Gradient Descent (SGD) and updated for better feature learning by CNN. At the last stage, the Multi-class LDA classifier evaluates the result into categorized classes. Based on the prediction, the tutors can easily find the low ranks of students who need a high preference for improving their academic performance. Experiments showed that the proposed model achieves greater accuracy (96.5%), precision (094), recall (092), F-score (095), and requires less computation time than existing methods.

Keywords: Multi-class LDA, CNN, mRMR, SGD, subjective pattern sequence.

*Corresponding author
Email addresses: srasheedmansoor@gmail.com (Rasheed Mansoor Ali S), perumal.scs@velsuniv.ac.in (S. Perumal)

Received: April 2021 *Accepted:* August 2021

1. Introduction

COVID-19 will negatively impact many parts of the world, as it has spread rapidly. It will have caused a serious ordeal for the human population. There were many alternatives for commercial activities and public functions discovered immediately during the lockdowns. The global economy has been severely impacted by the closure of entire businesses and travel restrictions, which have drastically changed people's lives all around the world. COVID-19 has had an impact on all degrees and types of training in higher education. During the recent economic downturn, a generation of young students has been forced to complete education uniquely. New factors and rules have emerged, influencing their ability to complete the current level of their schooling. Students who are exposed to Coronavirus face a range of challenges, including their education and professional development. These alterations have appeared in varying degrees around the world, depending on country-specific variables. Considering the pandemic's significance, it is essential to analyze the changes observed in teaching-learning examinations and their impact on university training. To increase students' interest in learning, it is necessary to measure students' academic and behavioral performance during this period. Data mining is used to extract vast amounts of information from unstructured and distributed data to analyze student behavior and interest in learning. Similarly, it assists classical educators in analyzing students' abilities and learning methods most effective for students.
Statistical, computer science, and machine learning technologies provide insight into learning from educational data, allow for a better understanding of learning performance, and enable effective teaching. Universities are now integrating digitalization in teaching-learning and other academic procedures, resulting in a massive volume of digital data. If this data is correctly transformed, it can assist teachers, policymakers, and administrators in making decisions. It improves the quality of educational processes by giving timely information to various stakeholders. Academic institutions attempt to develop a student model that can predict both the features and performance of each student. This will be highly supportive for lectures to support the low-ranking students effectively. Analyzing the collected data allows comparison with results from other test data and tracking dynamic changes over time.
The main objective of this analysis is as follows: 1. determine students' behavior towards learning 2. The learning characteristic and behavior of the student is collected 3. To analyze the data in a systematic way, formulate a conceptual framework; 4. Analyze the student data and uncover hidden relationships; 5. identify obstacles to distance learning, the potential changes, and future student learning improvement; 6. make recommendations for supporting the low-rank student. This paper develops a conceptual model for evaluating, comparing, and predicting student attitudes toward the education during COVID-19 crisis and provides a way to improve the low-ranking student. This data exploration reference proposed model enables students' perspectives and preparation for distant learning in an electronic world to be systematically assessed. Early discovery of problems could save students' education, as well as prevent certain negative social and economic effects on world growth. Several different fields can benefit from deep learning, including pattern recognition, image processing, object detection, and natural language processing. We utilized deep learning for building the proposed model to predict the student performance and behavior towards the learning. In this research, pre-processing techniques (like removing noise, and redundant data) are used to improve the accuracy of the results. The purpose of this paper, to construct a neural network with hidden layers and variable nodes based on new features and their weights. After building the system, these features and their weights are used to predict student information.
The rest of this paper is organized as follows, In section 2, we review previous research on students' academic performance during the COVID-19 pandemic. Section 3 introduces a new unified CNN-

Multi-class LDA framework for the evaluation of student learning performance and behavior. Section 4 Outlines an evaluation of the proposed framework on the real data. At last, the section 5 concludes and presents the future research plans.

2. Related Works

During the Covid-19 crisis, various research and analysis are being conducted to increase student performance in education. To predict student performance, researchers used a variety of techniques, including artificial neural networks, machine learning, and collaborative filtering. Li et al [7] proposed a paper on higher education students' performance prediction using deep learning approaches to help them choose courses and study schedules based on their skills. To evaluate the plan for education policy, the Adams and RMS prop improve the overall system performance. However, the algorithm has difficulty choosing the feature weight for classification. Li et al [8] proposed a deep learning framework to predict student performance in the course. SPDN (Sequential Prediction based on Deep Network) uses the multi-source fusion CNN technique to predict students' online behavioral sequences and includes static information via bidirectional LSTM. Meanwhile, students' internet usage has a greater influence on their academic success. Though it predicts more student data we can use it for some course prediction only.

Khattar et al [5], proposed a paper on the effect of covid-19 on student learning style. They took surveys with students about their studies and online classes, the responders are comfortable with the personal interaction with their friends than the peer group meeting online. By virtue of COVID-19, students face unpredictability regarding their grades and the path they will take by way of internships and jobs. It is a major factor in causing mental stress. The results also confirm that online teaching merely supplements classroom instruction and cannot replace the face-to-face interactions that occur in the classroom. Galina et al [3], proposed a basic framework for analyzing the effect of Covid-19 on the student's learning. Some of student cannot attend online classes due to a range of obstacles. The lack of an Internet connection or an electronic method of learning makes distance learning more challenging for low-income students. During this health crisis, poverty exacerbates the challenge of the digitization of education. An online survey was conducted by Minghat et al. in which 136 students from several universities in Indonesia and Malaysia were asked about the use of e-learning technologies during the COVID-19 outbreak in 2020. E-learning has had a positive influence and is now used by lecturers and students as an alternative learning method [10].

Lossoued et al [6] examined the challenges in obtaining quality online learning at the COVID-19 outbreak. A sample of 400 teachers and students' responses to a questionnaire were used in an exploratory descriptive technique by the researchers. According to the findings, teachers and students have faced self-imposed as well as pedagogical, technical, budgetary, and organizational barriers. Al-Okaily et al [1], the University of Jordan students have experienced a variety of environmental, electronic, and mental challenges owing to COVID-19. Over 220,000 Jordanian university students took part in COVID-19 through an online survey using university websites and portals. Botao et al [13], proposed a paper to classify the accident narratives in construction using the CNN model. It provides vital knowledge to the manager for improving the safety on-site.

Togascar et al [11], proposed a paper on detecting pneumonia by combining mRMR and machine learning models. The CNN deep features are applied to DT, kNN, LDA, LR, and SVM machine learning models. According to the findings of this study, deep features provided strong and consistent characteristics for pneumonia identification, and the mRMR technique improved classification efficiency. Junwen et al [2] proposed a paper on breast cancer prediction based on the mRMR algorithm which improves the feature selection process. Banghua et al [12], proposed a paper on improving the

accuracy of EEG motor imagery signal classification through LDA. By preserving the CSP spatial feature as well as using parameters for regulating the proportion of CSP spatial features in the training and test data sets, the LDA classifier's threshold has been updated. Khan et al. [4] published a paper on the estimation of student academic performance using a Bidirectional Long Short-Term Memory network (BiLSTM).
Mahareek et al [9] implement the simulated annealing algorithm to predict the student performance using SVM with Multilayer perceptron kernel (MLP kernel). Furthermore, several researchers have attempted to enhance student performance prediction using various deep learning approaches, but have yet to achieve a higher accuracy rate. To achieve a high rate of accuracy, we proposed the CNN-Multi-class LDA method for student performance prediction.

3. Proposed System

We describe our proposed CNN-multi-class LDA method for predicting student performance and behaviour in academics in this section. These are categorized into four categories: Dataset Acquisition, Pre-processing, Feature Selection, and Classification.

3.1. Acquisition of dataset

Students' performance analysis data has been acquired from the Kaggle repository [14]. For allocating test and train data, a dataset is split into two divisions with a ratio of 70:30.

3.2. Data Pre-processing

Developing an effective deep learning prediction model requires data preprocessing since raw data slows down classifier performance. The student prediction analysis dataset is available in a skewed format that contains numerous redundant data. It is essential to remove those redundant data from it and build a clean input for the machine learning model. We performed the following Pre-processing steps:

- Initially, noise is removed from the raw data
- Then redundant information is removed from it
- The categorical variables are transformed into a numeric form where certain attributes contain multiple values for single variables.

3.3. Feature Extraction

To calculate the frequency level, CNN extract the features from the pre-processed student log, calculate the interest rate for each subject. After the feature extraction, student interest and habitual features for different patterns are computed and sorted.
Algorithm: Feature Extraction based on Student interest
Input: Student Logs S_l
Output: Feature Weight F_w
Step 1: Student interest analysis S_{ia}

Compute frequency feature access

$$F_f = \frac{\sum S_{ia}.non\ like == feature\ analysis}{Total\ feature\ access\ +\ dterm\ repeated\ query}$$

Step 2: Compute student computed access weight.

$$S_{cm} = \frac{\sum S_{ia}.service==features}{total\ \ feature\ \ accessed} * S_{ia}(computed\ access\ weightage)$$

Step 3: compute relevance weightage score

$$If\ S_{cm} > S_c - Then\ rank\ list\ (service\ +\ behavioral\ interest\ of\ user)$$

Service interest score $\{S_{c1}, S_{c2}, ...\}$

Step 4: Return term of student featured rank list based on feature weight F_w
The above algorithm evaluates the interest features of each student for different performance patterns. The student performance and habitual patterns are analyzed and extracted the features from it.

3.4. Feature Selection

The mRMR algorithm reduces calculation costs by selecting the best features to be computed. Based on the variables' highest correlations, these features are chosen based on their similarity (equation (3.1)), which is determined by the variables' correlations.

$$I(M,N) = \sum_{m\in M}\sum_{n\in N}(m,n)\log(\frac{p(m,n)}{p_1(m)p_2(n)}) \tag{3.1}$$

The mRMR method's main goal is to pick the greatest traits and dilute the ones that aren't as good. Using the $I(M,N)$: measure the amount of similarity between the two features, M and N, this technique treats each feature separately from the dataset and uses the mutual information between them and measure the similarity of the features, M and N. For random variable, the marginal distribution function is calculated by combining the M and N probability.
Each probability function is defined as a vector in k-size $f_i = [f_i^1, f_i^2, f_i^3, ..., f_i^k$. The selected features are represented as S and class labels are denoted as H. The equation (3.2) and (3.3) has to be satisfied to obtain the best features,

$$Max\ D, D = \frac{1}{\mid S \mid}\sum_{f_i\in S} I(f_i, H) \tag{3.2}$$

f_i represent the mutual information among the variables where S set the maximum fit criteria based on that the unwanted variables are removed. Equation (3.3) minimizes the high rate of dependent variables and redundancy between the variables. Equations (3.2) and (3.3) calculate the maximum relevance ($Max\ D, D$) and minimum redundancy ($Min\ R, R$). Using equations (3.4) and (3.5), the optimization takes place at the same time.

$$MinR, R = \frac{1}{\mid S \mid^2}\sum_{f_i f_j\in S} I(f_i, f_j) \tag{3.3}$$

$$\max(D,R) = D - R \tag{3.4}$$

$$\max(D,R) = D/R \tag{3.5}$$

The SGD method updates the feature weight at every stage till the CNN learning process is optimized. The learning coefficient rate increase for CNN and provide the best features for mRMR to select which loop the process till the selection of the best features from the feature vector.

3.5. Classification

LDA (linear discriminant analysis) is a well-known technique for reducing and classifying dimensions. It's widely used in machine learning and pattern recognition, and it's shown to be effective in a variety of situations. By optimizing the scatter matrix trace between classes, and minimizing the scatter matrix trace within classes, the LDA determines the best projection vectors. It is a simple, powerful linear classification algorithm that can handle problems involving two classes or binary classification. Hence Multi-class LDA is preferred which can handle an arbitrary number of classes [15]. Let $x_1, x_2, ..., x_n$ be as samples and $y, y_2, ..., y_n$ are labels.

$$W_m = \sum_{i=1}^{n} n_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{3.6}$$

μ_k-over all sample mean of input data
k-class
n-number of classes
In addition, it computes statistics such as class means, scatters matrices, etc. Then it classifies the student performance into low, medium, or high.

3.6. Proposed CNN-Multi-class LDA method

The architecture diagram of the proposed CNN-Multi-class LDA method is depicted in Figure 1. The training and testing categories are first separated from the student prediction dataset. Then the training student log is applied to pre-processing where the noise and redundancy removal takes place. The CNN extracts the features from input student data by applying filters/kernel thereby generate the feature maps. After that, an activation function (ReLu) is passed over the output to provide a non-linear relationship. To avoid shrinkage of the feature map, paddings surround it and dimensional complexity is reduced by adding layers between the convolution layer.
The maximum pooling layer is utilized which sends maximum data to the next layer. Lastly, the vector data is flattened by the flattening layer. Features are extracted from the student log and converted into feature weights based on student interest and subject patterns. Based on the feature ranking it is sorted into a list and subjected to the feature selection process. It is imperative that the best features are selected from the feature map using the mRMR algorithm.

SDG updates the convolution layer based on new features on the layers. The Multi-class LDA classifies the results into three categories, such as low, medium, and high. Based on the prediction the low-ranking students will get high preference in their academic area.

4. Experimental Result

The proposed CNN-Multi-class LDA method is evaluated by a series of experiments using the Kaggle student performance analysis dataset. This model uses Python language to implement multi-class CNN-based student performance prediction. The Kaggle dataset is split into two parts with 70% of it being a training set and 30% of it being a test set. The value of the training parameters is used to evaluate its performance which is tabulated in Table 1. In the simulation environment, the models were compiled with graphics processing unit (GPU) support, and Windows 10 64-bit was used for the running of the operating system. Accuracy (ACC), Sensitivity (Se), Specificity (Sp), Precision (Pr), and F-score metrics are used to assess the models' performance. The proposed CNN-Multi-class LDA achieves 96.5% accuracy which is far better than the existing approaches SPDN (73.51%), Attention-based BiLSTM (90.16%), SVM (90.72 %) which is depicted in Figure 2.

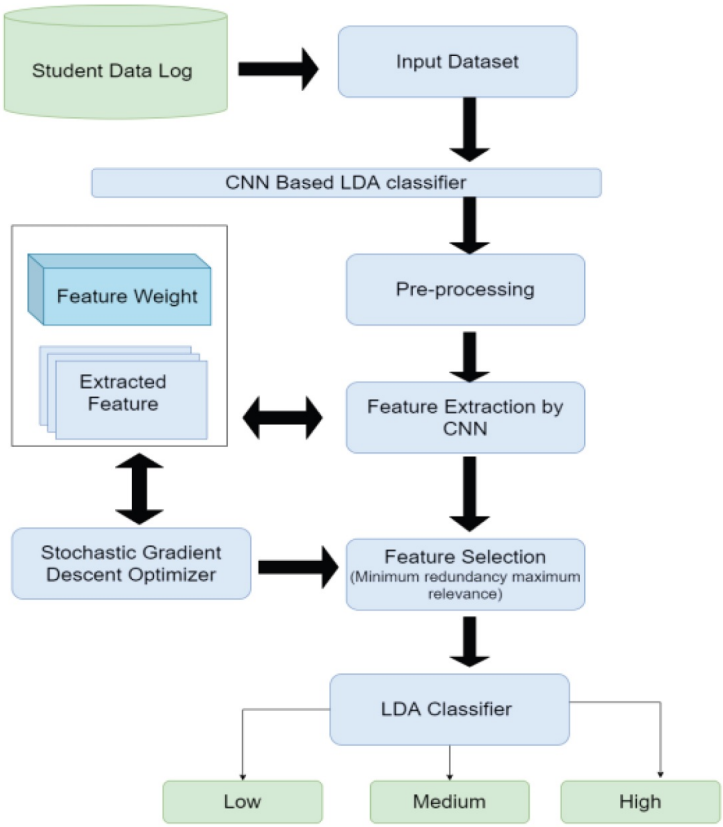

Figure 1: Architecture Diagram of proposed CNN-Multi-class LDA system

Table 1: Evaluating the proposed CNN- multi-class LDA performance in terms of accuracy, precision, recall and F-Score

Method	Accuracy	Precision	Recall	F-score
SPDN [8]	73.51	0.72	0.7	0.69
Attention-based BiLSTM [4]	90.16	0.9	0.9	0.9
SVM(MLP kernel) [9]	90.72	0.9	0.91	0.91
Proposed CNN-Multi-class LDA	96.5	0.94	0.92	0.95

The precision of the proposed CNN-Multi-class LDA model (0.94) is superior than the existing approaches SPDN (0.72), Attention-based BiLSTM (0.9), SVM (0.9) which is depicted in Figure 3. The Recall of the proposed CNN-Multi-class LDA model (0.92) is better than the existing approaches SPDN (0.7), Attention-based BiLSTM (0.9), SVM (0.91) which is depicted in Figure 4.

The F-score of the proposed CNN-Multi-class LDA model (0.95) is better than the existing approaches SPDN (0.69), Attention-based BiLSTM (0.9), SVM (0.91) which is depicted in Figure 5.

We evaluated the time complexity of the proposed CNN-Multi-class LDA method for different records (500,1000,1500,2000,2500) in Table 2 and Figure 6. It founded that the proposed method's computational time is lesser than the existing methods. With 500 records, the proposed method took only 3.2 seconds to execute in comparison with those existing methods (SPDN (3,9), Attention-based BiLSTM (3,5), and SVM (3,42). Similarly, the 2500 record took only 6.5 sec to compare to the existing methods (SPDN (7.5), Attention-based BiLSTM (7.3), and SVM(6.8)). CNN-Multi-class LDA method will provide significant improvements in the accuracy of student performance prediction, which will enable tutors to identify students' interests early and predict their behavior

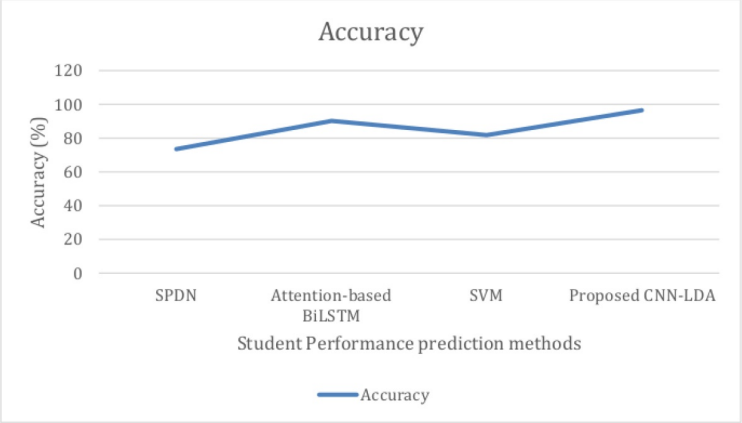

Figure 2: Accuracy of CNN- multi-class LDA method compared with the Existing methods (SPDN, Attention-based BiLSTM, and SVM)

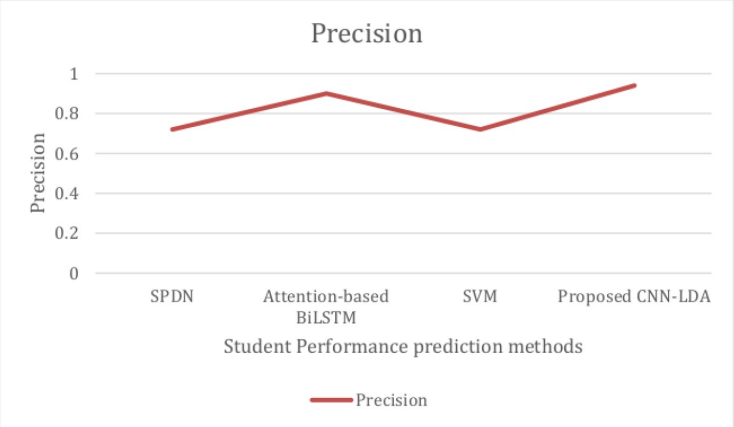

Figure 3: Precision of CNN- multi-class LDA method compared with the Existing methods (SPDN, Attention-based BiLSTM, and SVM)

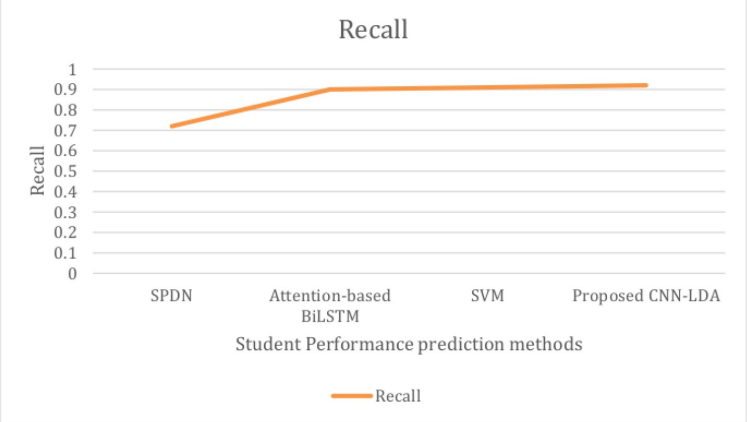

Figure 4: Recall of CNN- multi-class LDA method compared with the Existing methods (SPDN, Attention-based BiLSTM, and SVM)

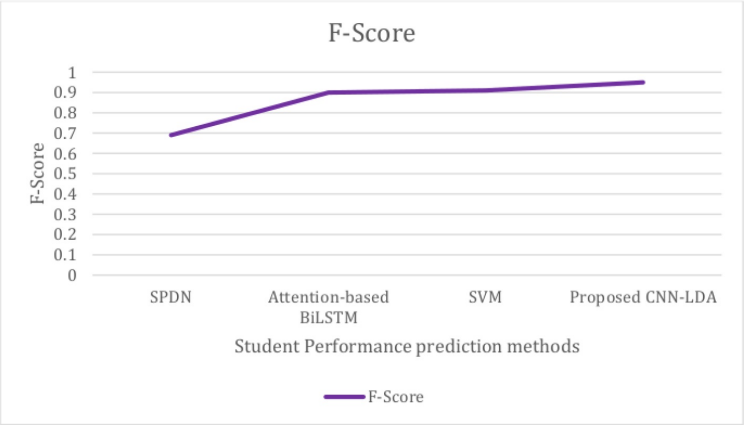

Figure 5: F-score of CNN- multi-class LDA method compared with the Existing methods (SPDN, Attention-based BiLSTM, and SVM)

Table 2: Time Complexity (sec) of proposed CNN-Multi-class-LDA

Number of records	SPDN [8]	Attention-based BiLSTM [4]	SVM(MLP kernel) [9]	Proposed CNN-Multi-class LDA
500	3.9	3.5	3.42	3.2
1000	4.8	4.7	4.5	4.1
1500	5.3	5.1	5	4.8
2000	6.5	6.2	5.6	5.2
2500	7.5	7.3	6.8	6.5

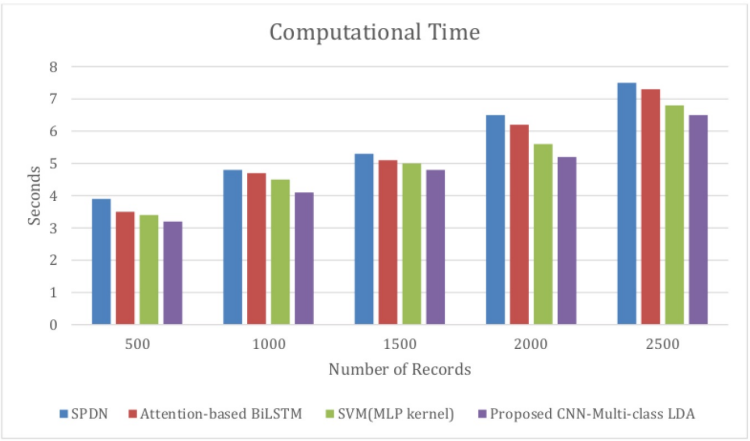

Figure 6: Computational Time of CNN- multi-class LDA method compared with the Existing methods (SPDN, Attention-based BiLSTM, and SVM)

toward academics. Thus, it will be a promising way to support the educational field in evaluating student progress and interest rates.

5. Conclusion

In this paper, we propose the CNN-Multiclass LDA method for predicting student performance and academic behavior. COVID-19 and the subsequent policy responses have caused profound changes in almost every aspect of our social and economic lives. To limit human mobility and disease spread, governments have enacted lockdowns and social distancing measures. The significant disruption of education during the COVID-19 crisis led to many schools switching from physical classrooms to online platforms to mitigate the learning losses associated with the disruptions. These disruptions widely affect the learning interest among the student and cause a declining result in academic performance. To solve this issues, predict an individual's interest rate and improve the performance of low-ranking students. We implement the CNN-based feature extraction which extracts valuable features from the student log then the best features are selected by the mRMR feature selection algorithm thereby eliminating the least features. The features are fetched to the Multi-class LDA for classification, which classifies the result into low, medium, and high. This prediction result supports numerous educational institutions and staff to identify low-ranking students. It also paves a way to motivate those students to achieve better performance in the future. For Future work, we are planning to increase the student data log and enhance the prediction efficiency.

References

[1] M. Al-Okaily, H. Alqudah, A. Matar, A. Lutfi and A. Taamneh, *Dataset on the acceptance of e-learning system among universities students under the COVID-19 pandemic conditions*, Data Brief 32 (2020) 106176.

[2] J. Di and Z. Shi, *Prediction model of breast cancer based on mRMR feature selection*, Inter. Conf. Neural Info. Proces. (2020) 32–40.

[3] G. Ilieva, T. Yankova, S. Klisarova-Belcheva and S. Ivanova, *Effects of COVID-19 pandemic on university students' learning*, Inf. 12(4) 2021.

[4] B. Khan, S.F. Khan, T. Rahman, I. Khan, I. Ullah, A.U. Rehman, M. Baz, H. Hamam and O. Cheikhrouhou, *Student-performulator: student academic performance using hybrid deep neural network*, Sustainability 13(17) (2021).

[5] A. Khattar, P.R. Jain and S.M.K. Quadri, *Effects of the disastrous pandemic COVID 19 on learning styles, activities and mental health of young Indian students - A machine learning approach*, 4th Inter. Conf. Intell. Comput. Cont. Syst. (2020) 1190–1195.

[6] Z. Lassoued, M. Alhendawi and R. Bashitialshaaer, *An exploratory study of the obstacles for achieving quality in distance learning during the COVID-19 pandemic*, Educ. Sci. 10(9) (2020) 232.

[7] S. Li and T. Liu, *Performance prediction for higher education students using deep learning*, Complexity 2021 (2021) 9958203.

[8] X. Li, X. Zhu, X. Zhu, Y. Ji and X. Tang, *Student academic performance prediction using deep multi-source behavior sequential network*, In: H. Lauw, R.W. Wong, A. Ntoulas, E.P. Lim, S.K. Ng and S. Pan (eds), Advances in Knowledge Discovery and Data Mining, PAKDD 2020, Lecture Notes in Computer Science, 2020.

[9] E.A. Mahareek, A.S. Desuky, H.A. El-Zhni, *Simulated annealing for SVM parameters optimization in student's performance prediction*, Bull. Elect. Engin. Inf. 10(3) (2021) 1211–1219.

[10] A.D. Minghat, A. Ana, P. Purnawarman, S. Saripudin, M. Muktiarni, V. Dwiyanti, S.S. Mustakim, *Students' perceptions of the twists and turns of e-learning in the midst of the Covid 19 outbreak*, Rev. Rom. Pentru Educ. Multid. 12(1Sup2) (2020) 15–26.

[11] M. Toğaçar, *A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models*, IRBM 41(4) (2020) 212–222.

[12] B. Yang, D. Li, B. Ma, X, Gu and D. Kong, *Motor imagery EEG classification method based on adaptive decision surface of LDA classifier*, 11th Int. Conf. Biosci. Biochem. Bioinf. Assoc. Comput. Machin. (2021) 37–41.

[13] B. Zhong, X. Pan, P.E.D. Love, L. Ding and W. Fang, *Deep learning and network analysis: classifying and visualizing accident narratives in construction*, Autom. Construc. 113 (2020) 103089.
[14] *https://www.kaggle.com/roshansharma/student-performance-analysis/data.*
[15] *https://multivariatestatsjl.readthedocs.io/en/latest/mclda.html.*