Business, Management and Economics Engineering
ISSN: 2669-2481 / eISSN: 2669-249X
2022 Volume 20 Issue 2

MULTISCALE ATTENTION-BASED DEEP LEARNING FOR STUDENTS' ACADEMIC GRADES PREDICTION

Dr. R. Parimaladevi
Assistant Professor, Department of Computer Science, PSG College of Arts & Science.

ABSTRACT

One of the most vital tasks in the domain of Educational Data Mining (EDM) is predicting students' academic success at an early stage of a semester. For this purpose, various Artificial Intelligence (AI) algorithms have been developed in the past years. Amongst, Deep Neural Network (DNN) can construct an effective predictive system by taking the academic features associated with the student's previous grades in the particular courses. On the other hand, it was unable to capture the multiscale temporal features of the time series data for students' success. Therefore, this article presents a new Multi-Scale Attention Deep Convolutional Neural Network (MSA-DCNN) to predict students' performance by extracting the academic and temporal features at different scales. Initially, the raw academic dataset is pre-processed by using different methods to convert it into the proper format. Then, the obtained dataset is fed to the MSA-DCNN, which applies the multiscale convolution to extract the academic and temporal features at multiple scales along with the occasion by creating various scales of feature maps. Also, an attention strategy is employed to decide relevant feature maps and reduce the feature dimensionality by learning the significance of all feature maps in an automated way. Further, the softmax classifier is used to predict the performance of students studying in a specific year. Finally, the experimental results illustrate that the MSA-DCNN achieves 94.9% of accuracy than the other state-of-the-art prediction models.

Keywords—Educational data mining, Artificial Intelligence, DNN, Academic features, Time series data, Temporal features, Multi-scale attention strategy, DCNN

I. INTRODUCTION

Education is critical to the advancement of a nation. It is also an important instrument for achieving success in life. Any academic institution strives to give its learners a high-quality education to improve the learning process. Students' academic achievement is a critical component that determines the success of any educational institution. Failure rates and dropout rates in computer programming courses are two major issues that students experience during the learning process at various levels of education [1-3]. Many AI models involving Machine Learning (ML) and Deep Learning (DL) algorithms have been used in a variety of applications like image analysis, language translation, speech recognition, text classification and EDM.

EDM is one of the most often adopted approaches for developing predictive models to extract hidden patterns and important information that may aid in education and learning [4]. It is concerned with the application of various data mining techniques in the education field such as classification, regression, time series analysis and association rule mining, to analyze various aspects of educational datasets collected from various e-learning environments or higher educational institutions. Educational institutions have begun to use AI technology to improve

students' learning processes. Today, educational institutions have significant difficulty in offering high-quality education to their students while also increasing their success rate [5-7]. In the area of education, ML and DL play a vital role in forecasting students' academic success in the future and assisting students in achieving higher grades. It is critical to forecasting students' academic achievement since it is a critical procedure to identify students who are in danger of failing at an early stage of a semester evaluation. As a result, these students will receive considerable retraining to improve their academic performance before the final evaluation and therefore raise the university's success rate [8].

Different data mining approaches were used to forecast diverse quality of education, including ability, accomplishment, attendance, dropout rate, and excellence [9-10]. Data mining methods are immensely useful in the subject of education, particularly for assessing and forecasting students' academic performance. Predicting students' academic performance at an early stage of a semester is a highly valuable technique for taking early interventions to improve their performance and also to lower student failure rates after a semester. On the other hand, predicting students' academic success is a difficult task since several elements might influence student performance, such as academic background, which is the prior academic achievements, demographic traits, economic background, behavioral features and other things. As a result, EDM is a crucial technique for addressing this issue [11].

One of the most popular applications of EDM is the use of previous academic data from students to forecast their future performance. It is a necessary tool that may be utilized to improve student performance, decrease failure rates and give full data on students' learning processes [12]. Typically, academic institutions generate large volumes of educational data, which are utilized for data analytics in the decision-making process to improve student performance. This might lead to a better understanding of the training process and an increase in overall educational environments [13-14]. The unbalanced dataset problem is one of the most significant problems impacting classifier performance. It is a major issue that arises in the area of EDM, resulting in inaccurate findings and inadequate training.

To deal with unbalanced classes, many resampling strategies such as SMOTE, ROS, ADASYN and SMOTE-ENN have been developed, which attains reliable outcomes and enhances the efficiency of predicting student’s performance. From these perspectives, Nabil et al. [15] developed the predictive systems based on the Deep Neural Network (DNN), Decision Tree (DT), Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbor (KNN), Random Forest (RF) and Gradient Boosting (GB) to assess the database gathered from an open 4-year university from 2006 to 2020. Also, it was used to estimate students’ success in a Data Structure course and detect at-risk students at an early stage of a semester depending on their grades in the previous courses of the initial academic year. The major processes involved in this system were data acquisition, pre-processing, modeling and prediction. Although DNN outperforms other ML algorithms, it considers only academic features related to the student’s grades in the courses of the initial academic year to construct the predictive system, whereas the multiscale temporal features extracted from the time series data were essential for the prediction task. But, the DNN may not fit for capturing the temporal features in the time series academic data.

Hence in this paper, the MSA-DCNN framework is proposed to enhance the efficiency of student's performance prediction task. First, the pre-processed academic dataset is given to the MSA-DCNN classifier for prediction. In this MSA-DCNN framework, a stacked multi-scale attention unit is designed to perform the multi-scale attention strategy that captures the most significant feature maps and discards less significant features by automatically learning the significance of all feature maps. The multi-scale attention unit has a multi-scale block and an attention block. The multi-scale block is used to create various dimensions of the receptive field which extracts the multiple scales of features by the fused convolutional layer with different kernel size. Similarly, the attention block is used to obtain the significant feature maps based on the compressing and rescaling mechanisms. Moreover, the softmax function is applied to predict the students' performance in a certain period. Thus, this MSA-DCNN can predict the student's performance with the aid of both academic and temporal features at different scales effectively.

The rest of the paper is prepared as the following: Section II discusses the work associated with the prediction of student's performance. Section III explains the design of proposed algorithm and Section IV portrays their efficiency. Section V summarizes this paper and suggests future improvements.

II. LITERATURE SURVEY

Kim et al. [16] developed a novel deep learning-based model called GritNet, which constructs based on the Bi-directional Long Short-Term Memory (BLSTM) and GMP layers. In this model, the student's raw event records were given as input to the GridNet to encode the time-stamped logs into a sequence of fixed length input vectors. But, it needs indirect data such as student board activity and relationships with a mentor were needed to increase the prediction efficiency.

Francis and Babu [17] designed a novel prediction method to analyze students' performance in academia based on the classification and clustering algorithms. First, the student data was pre-processed and various features were extracted. Then, these features were classified by the Support Vector Machine (SVM), naive Bayes, decision tree and neural network classifiers to predict which of the features provide higher accuracy. After that, such features were fed to the K-means clustering and applied a majority voting scheme to predict students' academic performance. But, it does not handle huge varieties of features from the student database.

Li et al. [18] suggested the fuzzy C-means clustering algorithm to predict the student's performance according to their examination results. However, it considers only the student's grades while other characteristics of the students were essential to enhance the prediction accuracy. Hassan et al. [19] developed an ensemble classification model using different machine learning algorithms to predict the student's performance and recognize at-risk students from their earlier stage. But, this model was time-consuming for large-scale datasets.

Shao et al. [20] designed an optimized mining algorithm to analyze students' learning degrees based on the dynamic data. Initially, the optimized text classification was used to map the query texts to the knowledge points automatically. After, the subjective weighting scheme was integrated with the expert knowledge to produce the training level matrix of students on knowledge records. Based on this matrix, the DBSCAN clustering was used to group the personalized learning characteristics of students. But, it depends on expert knowledge and considers only dynamic data.

Hooshyar et al. [21] developed an improved technique based on the student's assignment submission behavior to predict their activities with learning complexities via procrastination behavior. First, the feature vectors were constructed for defining the submission behavior of students for every assignment. After, clustering was performed for labeling students as a procrastinator, procrastination candidates or non-procrastinator. But, the length of the feature vectors was predetermined.

Rodríguez-Hernández et al. [22] utilized Artificial Neural Network (ANN) to classify students' academic efficiency as either high or low. But, the significant data defined by the high school grade point average was not considered. Also, each data regarding students' socioeconomic situations was self-reported by the students, which tends to be an inaccurate prediction. Matzavela and Alepis [23] developed a decision tree learning using a predictive framework to classify the student's knowledge level based on the weights of the decision tree. But, it needs more dynamic features that provide an effective m-learning platform in tertiary education.

Giannakas et al. [24] suggested the DNN model with 2 hidden layers to predict the team's performance in the software engineering course. But, it does not consider the time-series data and the training parameters were not optimized efficiently. Dabhade et al. [25] suggested the linear regression and support vector regression algorithms to determine the academic performance of the final year undergraduate students of the particular institutions. But, these algorithms were prone to missing data and not suitable for a large number of features extracted from the student database.

III. PROPOSED METHODOLOGY

In this section, the proposed MSA-DCNN framework for predicting students' performance is described briefly. Figure 1 depicts the flow diagram of the presented predictive system to predict the students' performance using MSA-DCNN framework. This system consists of the following phases: data acquisition, data pre-processing, training and testing.

3.1 Data Acquisition and Pre-Processing

First, a dataset is acquired which contains different records of anonymized students with various features about their past academic achievements during the initial academic years. Then, different pre-processing methods are performed including data cleaning, discretization, feature encoding, imbalanced datasets handling and data scaling to transform the raw data into the appropriate form during prediction [15].

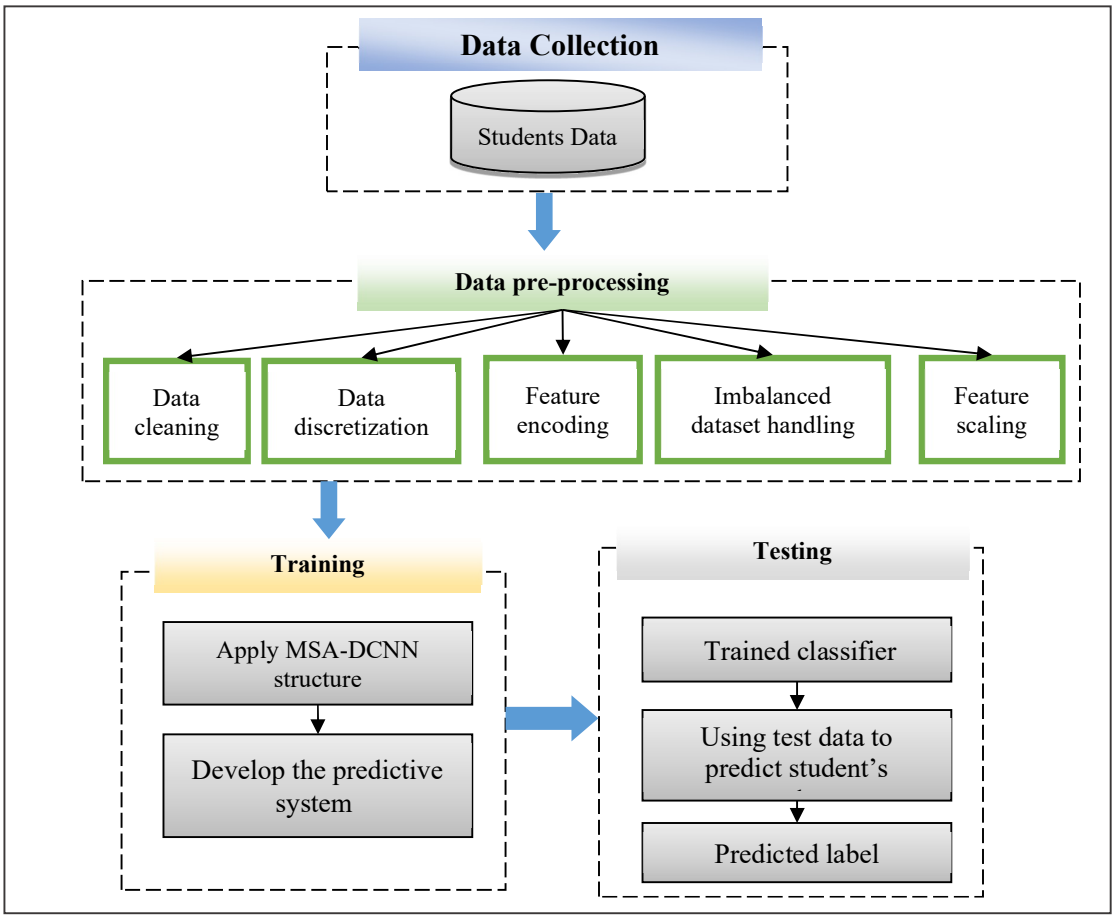

Figure 1. Flow Diagram of the Proposed Student's Performance Predictive System

3.2 Prediction using MSA-DCNN Classifier

The pre-processed dataset is fed to the MSA-DCNN structure which is portrayed in Figure 2. The symbols m, p and s represent the MSA, pooling and softmax layer, correspondingly. The symbols L and N are the time series length and the number of classes, correspondingly. Observe that the numbers before and after × indicate the length and channels of feature maps, respectively.

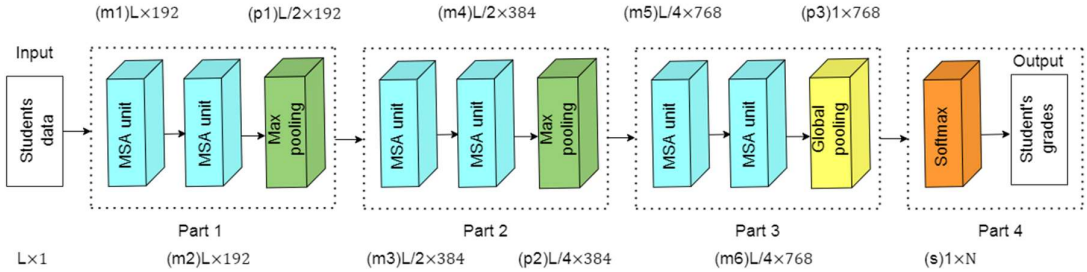

Figure 2. Entire Architecture of MSA-DCNN Framework

In MSA-DCNN, four major parts are involved. For the primary 2 parts, each unit comprises the three different layers. The initial 2 layers are the multiscale layer, which is the core layer of the entire structure. The MSA layer has a MSA unit that executes the Multiscale Attention Strategy (MAS) to learn the significance of all feature maps created by the multiscale convolution. The final layer is the max-pooling layer which decreases the number of variables

to avoid overfitting and enhance the model efficiency by optimizing the features with robust prediction.

The architecture of the third part is nearly similar as the past parts excluding the pooling layer. In this part, a global mean pooling layer is applied rather than the Fully Connected (FC) layer to create the final feature maps for prediction. The final part is the output unit, which offers the prediction outcome by the softmax layer. The softmax layer offers the posterior likelihood of all classes which is coded as one-hot form and passed to the output.

Multiscale Attention Strategy

The MAS is an approach that improves valuable feature maps while suppressing less useful ones based on the significance of all feature maps created by multi-scale convolution. The MAS's aim is to increase a network's recognizing capabilities. To accomplish this aim, multi-scale convolution is recommended initially to extract multi-scale temporal data, which relates to short-term, mid-term and long-term correlations of time series. After that, the attention strategy is used to decide the most relevant features while ignoring the unnecessary ones by rescaling the weight of all feature maps. As a result, MSA is projected to boost prediction efficiency.

The MAS is executed by the MSA unit which is the major unit of the MSA-DCNN structure. The architecture of MSA unit has a multiscale block and an attention block as illustrated in Figure 3. The symbols C_1, L_f and C_2 are the channels of input feature maps, the length of input feature maps and the output channels of all convolutional layer in the multiscale block, correspondingly.

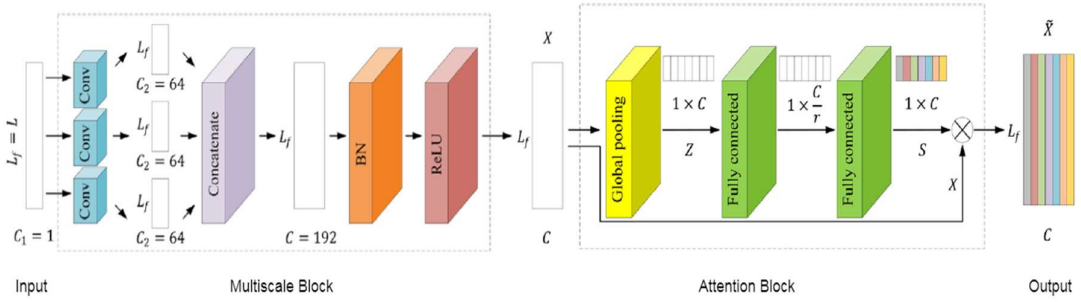

Figure 3. Architecture of MSA Unit

Consider $X \in \mathbb{R}^{L_f \times C}$ and $\tilde{X} \in \mathbb{R}^{L_f \times C}$ are the outcome of the multi-scale block and the attention block, correspondingly, where C is the number of channels and observed that $C = 3C_2$ because there are 3 convolutional layers in the multi-scale block is constructed by the fused convolutional layers with various kernel sizes. The multiscale block creates various dimensions of the receptive field to obtain various scales of temporal data whereas the attention unit fetches in a channel-wise weight to focus on the significant feature maps that improves the prediction ability. Observe that, because the input is the raw time series, the channels of input feature maps are 1 and the length of input feature maps is L. Also, the number of filters is assigned to 64 in the convolutional layer of the initial MSA unit. So, the output channels of all convolutional layers are 64.

The architecture of the multiscale block involves the four different layers. The first layer is the convolutional layer which employs a convolution function to the input data with shared weights to capture the features and create the feature maps. In this framework, 3 paralleled

convolutional layers are employed as the multiscale convolution to learn the feature maps of time series, which is intended at extracting various temporal features. The key benefit of this multiscale convolution is that it generates various dimensions of the receptive field to extract a vast amount of information from the previous layer than the single-scale convolution.

The second layer is the concatenation layer, which merges the feature maps of various convolutional layers on the dimension of channels. The third layer is the batch regularization layer, which preserves similar input distribution during the network learning by regularizing the feature maps of preceding layers. In this manner, the gradient vanishing problem is solved by a greater gradient, which tends to high convergence speed during training phase.

The final layer is the rectification layer with ReLU as the activation function, which enhances the nonlinear correlation among the layers and creates the sparsity to avoid overfitting issue. It allows robust training to predict the performance of students.

The architecture of the attention block comprises the 3 different layers. The initial layer is the global mean pooling layer, which compresses the global temporal data into the channel descriptor to create channel-wise data by determining the average of each values for all feature maps. The input of the attention block is the outcome of the multiscale block X. Consider $Z \in \mathbb{R}^C$ is the channel-wise data, which is created by reducing X via the feature map length L_f. Then, the n^{th} element of Z is determined as:

$$z_n = \frac{1}{L_f}\sum_{i=1}^{L_f} x_n(i) \quad (1)$$

The final 2 layers are FC layers, which accomplish the aim of dynamic recalibration to obtain the channel-wise weight S by decreasing the feature map channels with a minimization factor r and then rescaling it to the actual dimension. Particularly, the channels of the initial FC layer with an activation function of ReLU is $\frac{C}{r}$ and the channels of the second FC layer with an activation function of sigmoid is C, where the n^{th} element of S is acquired as:

$$s_n = \sigma\big(W_2\delta(W_1 z_n)\big) \quad (2)$$

In Eq. (2), δ and σ are the ReLU and sigmoid activation function, correspondingly, $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are variables of the 2 FC layers, correspondingly. The outcome of the attention block $\tilde{X}$ is acquired by employing the channel-wise weight S to the feature maps, where the n^{th} element of $\tilde{X}$ is determined as:

$$\tilde{x}_n = s_n \cdot x_n \quad (3)$$

Thus, the academic and temporal features are extracted by the MSA unit at different scales, which are then fed to the softmax layer to predict the students' grades in a certain year.

IV. EXPERIMENTAL RESULTS

In this section, the performance of MSA-DCNN is analyzed by implementing it in MATLAB 2017b and compared with the existing classification algorithms: DT, LR, SVC, KNN, RF, GB, DNN [15], ANN [22] and fuzzy C-means [18].

In this experiment, a total of 12054 data instances are collected from the different government and self-financed engineering colleges in the duration from Jan 2011 to Dec 2021. Amongst, 8438 data instances are used for training and 3616 data are used for testing. In the dataset, 39 attributes are present which includes students' name, age, gender, course, nature of college such as medical/engineering, college type like government, self-financed, location feature,

family belong to nuclear family or joint family, family factors such that occupation & educational qualification of family members, spending time in television, college factors, economic factors, social factors and, mobile, computer, etc., personal factors, academic factors. For instance, location features defined as the location in which students' home, school and college placed such as rural area, urban area and semi-urban area. College factors is known as the study materials like lecturer notes & book materials, method of teaching, number of students in class, allowance of mobile phones, etc. Social factors such that guidance of relatives for studies, number of friends and academic performance of friends.

The considered evaluation metrics to analyze the efficiency are described below:

- **Accuracy:** It is the percentage between a proper prediction of students' grades and the total number of predictions performed.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

  True Positive (TP) and True Negative (TN) are the solutions where classifier predicts the pass and fail grade classes as themselves, correspondingly. False Positive (FP) is a solution where classifier improperly predicts the fail grade classes as pass, whereas False Negative (FN) is a solution where classifier improperly predicts the pass grade classes as fail.

- **Precision:** It is the percentage of properly predicted the students' grade classes at TP and FP rates.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

- **Recall:** It is the percentage of properly predicted students' grade classes at TP and FN rates.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

- **F-measure:** It is calculated as the harmonic mean of precision and recall.

$$F - measure = 2 \times \frac{Precision \cdot Recall}{Precision+Recal} \quad (7)$$

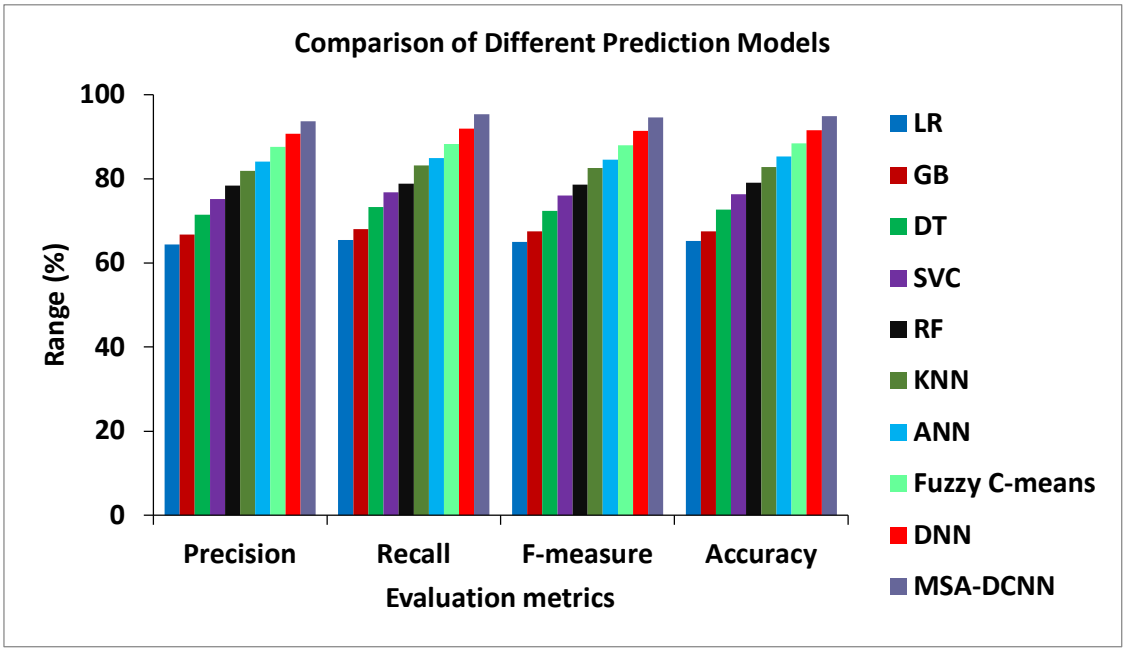

Figure 4. Performance Analysis of Different Students' Performance Prediction Models

Figure 4 illustrates the efficiency of various models applied to predict the students' grades. It observes that the effectiveness of the MSA-DCNN in terms of precision, recall, f-measure and accuracy is greater than that of other prediction models because of extracting both academic and temporal features at multiple scales from the student database. The precision values obtained by the LR, GB, DT, SVC, RF, KNN, ANN, fuzzy C-means, DNN and MSA-DCNN are 64.4%, 66.8%, 71.5%, 75.2%, 78.4%, 81.9%, 84.1%, 87.6%, 90.8% and 93.7%, correspondingly. The recall values determined by the LR, GB, DT, SVC, RF, KNN, ANN, fuzzy C-means, DNN and MSA-DCNN are 65.5%, 68.1%, 73.3%, 76.8%, 78.9%, 83.2%, 85%, 88.3%, 92% and 95.4%, correspondingly. Similarly, the f-measure values achieved by the LR, GB, DT, SVC, RF, KNN, ANN, fuzzy C-means, DNN and MSA-DCNN are 65%, 67.5%, 72.4%, 76%, 78.7%, 82.6%, 84.6%, 88%, 91.4% and 94.6%, correspondingly.

Accordingly, the accuracy of MSA-DCNN is 45.55% greater than the LR, 40.59% greater than the GB, 30.54% greater than the DT, 24.2% greater than the SVC, 19.97% greater than the RF, 14.6% greater than the KNN, 11.25% greater than the ANN, 7.23% greater than the fuzzy C-means and 3.6% greater than the DNN for predicting the students' grades and preventing their dropout rate.

V. CONCLUSION

In this study, the MSA-DCNN framework was presented for students' performance prediction with the aid of capturing the academic and temporal features at different scales. In this system, the raw academic database was pre-processed and provided to the MSA-DCNN for obtaining the academic and temporal features at multiple scales by using the multiscale convolution. Additionally, the attention policy was used to choose the most significant feature maps by learning the significance of all feature maps. Besides, the softmax classifier was employed to predict the students' performance efficiently. At last, the test outcomes realized that the MSA-DCNN has an accuracy of 94.9% compared to the other predictive models.

REFERENCES

[1] Ramaphosa, K. I. M., Zuva, T., & Kwuimi, R. (2018). Educational data mining to improve learner performance in Gauteng primary schools. In *IEEE International Conference on Advances in Big Data, Computing and Data Communication Systems*, pp. 1-6.

[2] Bennedsen, J., & Caspersen, M. E. (2019). Failure rates in introductory programming: 12 years later. *ACM Inroads*, *10*(2), 30-36.

[3] Buschetto Macarini, L. A., Cechinel, C., Batista Machado, M. F., Faria Culmant Ramos, V., & Munoz, R. (2019). Predicting students success in blended learning—evaluating different interactions inside learning management systems. *Applied Sciences*, *9*(24), 1-23.

[4] Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462-55470.

[5] Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering*, *928*(3), 1-18.

[6] Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2021). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 1-4.

[7] Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial Intelligence and Machine Learning to Predict Student Performance during the COVID-19. *Procedia Computer Science*, *184*, 835-840.

[8] Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1-21.

[9] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 1-27.

[10] Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of students performance using machine learning. In *IOP Conference Series: Materials Science and Engineering*, *1055*(1), 1-8.

[11] Khalaf, A., Dahr, J. M., Najim, I. A., Kamel, M. B., Hashim, A. S., Awadh, W. A., & Humadi, A. M. (2021). Supervised learning algorithms in educational data mining: a systematic review. *Southeast Europe Journal of Soft Computing*, *10*(1), 55-70.

[12] Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., ... & Weegar, R. (2021). Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. *Frontiers in Artificial Intelligence*, *4*, 1-20.

[13] Khan, I., Ahmad, A. R., Jabeur, N., & Mahdi, M. N. (2021). An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, *8*(1), 1-18.

[14] Baneres, D., Rodríguez-Gonzalez, M. E., & Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*, *12*(2), 249-263.

[15] Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, *9*, 140731-140746.

[16] Kim, B. H., Vizitei, E., & Ganapathi, V. (2018). GritNet: student performance prediction with deep learning. *arXiv preprint arXiv:1804.07405*.

[17] Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of medical systems*, *43*(6), 1-15.

[18] Li, Y., Gou, J., & Fan, Z. (2019). Educational data mining for students' performance based on fuzzy C-means clustering. *The Journal of Engineering*, *2019*(11), 8245-8250.

[19] Hassan, H., Ahmad, N. B., & Anuar, S. (2020). Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. In *Journal of Physics: Conference Series*, *1529*(5), 1-8.

[20] Shao, Z., Sun, H., Wang, X., & Sun, Z. (2020). An optimized mining algorithm for analyzing students' learning degree based on dynamic data. *IEEE Access*, *8*, 113543-113556

[21] Hooshyar, D., Pedaste, M., & Yang, Y. (2020). Mining educational data to predict students' performance through procrastination behavior. *Entropy*, *22*(1), 1-24.

[22] Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*, 1-14.

[23] Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. *Computers and Education: Artificial Intelligence*, *2*, 1-12.

[24] Giannakas, F., Troussas, C., Voyiatzis, I., & Sgouropoulou, C. (2021). A deep learning classification framework for early prediction of team-based academic performance. *Applied Soft Computing*, *106*, 1-17.

[25] Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, 1-8.