International Journal of Medical Toxicology & Legal Medicine Volume 27, No. 2, 2024
e-ISSN: 0974-4614
p-ISSN: 0972-0448

Predicting Student Academic Achievement Using Data Mining and Deep Learning Techniques in Educational and Medical-Legal Contexts

Dr Gowtham Mamidisetti[1], Praveen Talari[2], Dr. S. Suguna Mallika[3], Mrs. Banoth Rajeshwari[4], Sai Hemanth Maremalla[5], Kayam Saikumar[6]

[1]Professor, Department of Information Technology, Malla Reddy University, Maisammaguda, Dulapally, Hyderabad, Telangana 500043, Email: drmgowtham@mallareddyuniversity.ac.in
[2]Department of CSE, Vignana Bharathi institute of technology, Hyderabad, Telangana 501301 , Email: praveent738@gmail.com
[3]Professor of CSE, CVR College of Engineering, Telangana India, 501510, Email: ss.mallika@cvr.ac.in
[4]Department of Computer Science andEngineering, MLR Institute of Technology, Hyderabad,Telangana, India, Email: banothchina@mlrit.ac.in
[5]software engineer, university of alabama at Birmingham, Bhaskarapuram,machilipatnam, 521001 , Email: maremallasaihemanth@gmail.com
[6]Assistant Professor, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India , Email: saikumar.kayam@klh.edu.in

Received: 15.08.2024 Revised: 20.09.2024 Accepted: 22.10.2024

ABSTRACT
Educators are starting to take an interest in intelligent technology development. Conventional processing methods may be inadequate and skewed due to the exponential growth of educational data. Therefore, it is more crucial than ever to replicate data mining research technologies for use in the field of education. Using relevant theories of grouping, discriminating, and convolutional neural networks, this study assesses and forecasts students' academic performance in order to avoid inaccurate assessment results and to plan for their future performance. Using a statistic that has never been utilized in the K-means approach before to improve the clustering-number determination is the first advice from this study. After that, we'll assess the K-means method's clustering efficacy using discriminant analysis. A convolutional neural network, which can be trained and evaluated with labelled data, is shown. The produced model can be used to forecast future performance. The last step is to use two metrics in two cross validation methods to assess the constructed model and verify the predicted findings. The experimental results show that the statistic makes the results more predictable and answers the quantitative and objective question of how to get the clustering number using the K-means method. The accuracy of 98.34%, recall of 97.85%

Keywords: Academic performance, clustering analysis, convolutional neural networks, discriminant analysis, educational data mining.

1. INTRODUCTION

The goal of data mining (DM) is to extract useful information from large datasets that lack any clear organisation. Data mining, machine learning, and statistical techniques are the centre of attention in the field of educational data mining (EDM). Investigations into the potential benefits of data mining tools for use in classrooms have been ongoing for some time. Given the proliferation of accessible data sets and AI-powered learning platforms, its profile has risen in recent years [1]. Engineering data mining (EDM) entails developing and deploying data mining algorithms to investigate massive datasets derived from a wide range of sources. A key component of any gooduniversity is the quality of its graduates' academic work.
Thus, EDM professionals consider it a top priority to foresee how students will learn and to evaluate their progress[2].

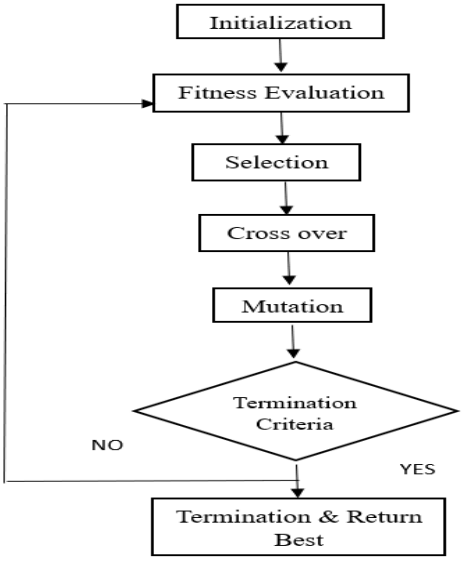

Fig.1: A sample figure for the data.

EDM's ever-expanding disciplinary concentration is on self-learning and adaptive algorithms that detect internal connections or patterns in educational data. As part of the big data paradigm, heterogeneous data is expanding its presence in the realm of education. Finding useful insights in massive educational datasets adaptively calls for certain data mining techniques [3]. The rapid advancement of EDM application research can be attributed to the fact that data mining technologies allow for the analysis of massive volumes of student data in order to uncover valuable patterns of student learning behaviour. Data mining tools have improved many areas of educational data processing, such as retaining students, predicting when they would drop out, analysing academic data, and studying student conduct [4]. Predicting and evaluating students' academic performance has always been a top priority for EDM.

2. LITERATURE REVIEW

A systematic review of deep learning approaches to educational data mining:

There is a mountain of data kept by schools nowadays, including students' exam results, attendance records, and enrolment information. Mining this kind of data produces exciting new information that its consumers will find beneficial. As educational data continues to expand at an exponential rate, a more sophisticated collection of algorithms will be needed to mine these enormous databases for actionable insights. In light of this problem, the area of educational data mining (EDM) was born. There is no clear way to address educational difficulties using traditional data mining approaches, even though they may serve a specific purpose. Thus, specialised data mining methodologies cannot be applied to the challenges without first implementing a pretreatment procedure. An example of an EDM preprocessing method is clustering. Numerous EDM studies have focused on data mining techniques applied to educational qualities. Therefore, this study supplies a comprehensive literature review covering the three decades (1983–2016) of clustering algorithms, their applications, and their value in EDM [5]. Future insights are offered and avenues for further research are suggested based on the literature review [6].

Implementing AutoML in educational data mining for prediction tasks:

Emerging within the last 20 years, Educational Data Mining (EDM) focuses on creating and implementing data mining techniques to facilitate the analysis of large amounts of data collected from various educational environments [7]. Predicting students' growth and learning outcomes, including absence, performance, and grade point average, is a crucial part of working in the EDM field. Consequently, for accurate prediction models, it is imperative that data scientists and educators employ suitable machine learning approaches [8]. Because of the high-dimensional input space and methods employed in machine learning, building accurate and reliable learning models may be a challenging and error-prone process that requires good data science talents. It might be difficult to understand and articulate the results of a study when the methods used to solve it are not intuitive or well-suited to the problem at hand [9]. Through hyperparameter optimisation, this work aims to explore the potential applications of advanced machine learning algorithms in educational settings. More specifically, we investigate how well automated Machine Learning (autoML) can forecast students' achievement in online courses in relation to their participation in those courses. Simultaneously, we limit to models that rely

on rules and trees so that the results can be seen and understood [10]. According to the results of the many experiments undertaken with this aim in mind, auto ML tools consistently produce superior results. The ultimate goal of our work is to make it easier for people who aren't experts in EDM, like teachers and professors, to do tests with the right automated parameter settings and get results that are both accurate and easy to understand [11].

Integration of data mining clustering approach in the personalized E-learning system:
A relatively new area of study, educational data mining aims to enhance flexible methods of learning and self-study. Finding underlying structures or patterns in educational data is its primary function. A growing part of the big data paradigm in education is heterogeneous data. Specialised data mining methods are required to adaptably retrieve valuable information from large datasets pertaining to schooling. Students' distinct patterns of behaviour in the classroom can be better understood using the clustering method detailed in this research. Also displayed is the architecture of the personalised e-learning system, which can detect the students' learning capacities and respond accordingly to course contents [12]. Finding the optimal settings where pupils can enhance their learning ability is the primary goal. In addition, the government can uncover significant hidden trends that will guide their efforts to improve the current system. Clustering techniques include K-Means and K-Medoids as well as Agglomerative Hierarchical Cluster Tree and Density-based Spatial Clustering of Applications with Noise, and Clustering by Fast Search and Finding of Density Peaks through Heat Diffusion (CFSFDP-HD) are examined through the lens of educational data mining. Results are more robust when CFSFDP-HD is used instead of existing methods [13]. Analysing large datasets to enhance educational systems is another equally valuable use of data mining tools [14].

The use of tools of data mining to decision making in engineering education—A systematic mapping study
Research on educational data mining, both theoretical and applied, has been on the rise in the past few years. The field of learning analytics makes use of methods, software, and algorithms to improve the quality of instruction by allowing users to find and extract meaningful patterns in previously collected student data [15]. However, learning analytics pays little attention to several requirements related to the integration of new technology into the pedagogical and scholastic processes [16]. Research on the topic of learning analytics in engineering education has not yet yielded a systematic evaluation. This article presents a study that summarises the work made so far and identifies areas that require further research. To do this, a thorough mapping study was carried out with the intention of classifying publications according to the type of research and the contribution they made [17]. Software and computer science engineers seem to be the target audience for case study research, according to the results. Also, new applications of learning analytics can be seen in areas like predicting whether a student will stay in school or drop out, analysing academic data, evaluating student learning, and analysing student behaviour [18]. While this systematic mapping study primarily focused on engineering education and learning analytics, its results might be useful in other fields as well [19].

Data mining in educational technology classroom research: Can it make a contribution?:
Concerning the use of data mining in the study of instructional technologies, this work addresses and elucidates a number of critical topics [20]. As case studies, two studies in Europe and Australia illustrate the use of data mining techniques, including fuzzy representations and association rules mining. Students' learning, behaviour, and experiences with computer-supported classroom activities are the focus of both research [21]. The first study used an association rules mining approach to learn about the many ways in which students' cognitive types interacted with a simulation to find a solution [22]. For reliable information about how students used and performed in the simulation, association rules mining proved to be an excellent tool. Findings from this study suggest ways in which data mining might enhance current practices for evaluating educational software. In the second study, researchers used fuzzy representations to inductively analyse survey responses [23]. Based on the findings, educational technologists can better plan and assess technology integration initiatives in schools by making use of data mining [24]. Implications for developing instructional data mining tools that effectively convey results, details, explanations, remarks, and recommendations to users who lack expertise in the field are discussed in light of this study's implications [25]. The handling of data privacy concerns is completed at last [26].

3.METHODOLOGY
There are limitations to the traditional absolute score when it comes to providing an accurate representation of the learning environment. Some of the reasons for This encompasses the reality that courses vary in terms of difficulty and that grading standards, even between instructors in the same class, can vary greatly. Colleges and universities should ensure talent quality by not just evaluating students based on grades, but also by studying the effects of students' learning, making predictions about students' future academic performance using these

estimates, and issuing timely academic warnings. The goal of this initiative is to improve educational resource management by helping both students and colleges and universities raise the bar on educational quality.

There are two main types of data mining methods: supervised and unsupervised. When making predictions, unsupervised algorithms scour datasets for patterns in the absence of knowledge about the output variables. Supervised methods, on the other hand, employ the inputs to forecast the values of the output variables; these approaches are also known as guided or predictive methods. Here, we'll be looking specifically at supervised DM algorithms, which estimate the values of output variables from inputs. In order to achieve this, training data is used to label input and output values, which are then used to train a model.

To depict a collection of variables and the conditions upon which they rely, Bayesian networks use probabilistic graphical models. They are constructed using directed acyclic graphs. When trying to determine the probability that a certain known cause contributed to an already-known event, they are ideal. An ADTree, short for Alternating Decision Tree, is one method for classification in machine learning. It provides resources for boosting and expands the use of decision trees. Nodes in this network alternate between describing predicate conditions and carrying single numbers. After counting all the paths with true decision nodes and true prediction nodes, an ADTree determines the instance's classification. Bayes Net enhances classification accuracy when utilised in combination with an ADTree.Examining students' academic performance from a clustering perspective and making predictions about their future performance based on their current performance is the central research question of this work.

A training set, which includes a database of past observations labelled with predetermined categories, is utilised in the Classification and Regression Tree (CART) study to construct decision trees. Depending on the type of data collection, this learning method might provide regression trees or classification trees as output. From the set of trees taken into account during pruning, it chooses the best one using cross-validation or a big independent test sample. During the CART implementation step, the dataset is divided into two subgroups based on the degree of dissimilarity in the results.

One approach to data classification that relies on learning a logical phrase is LAD. Amudha et al. (2011) noted that LAD can distinguish between positive and negative samples due to its status as a binary classifier. After LAD processes a dataset, it generates a huge number of patterns. To ensure that every pattern in the model meets certain criteria regarding homogeneity and prevalence, a subset of these patterns is chosen to meet the aforementioned assumption.

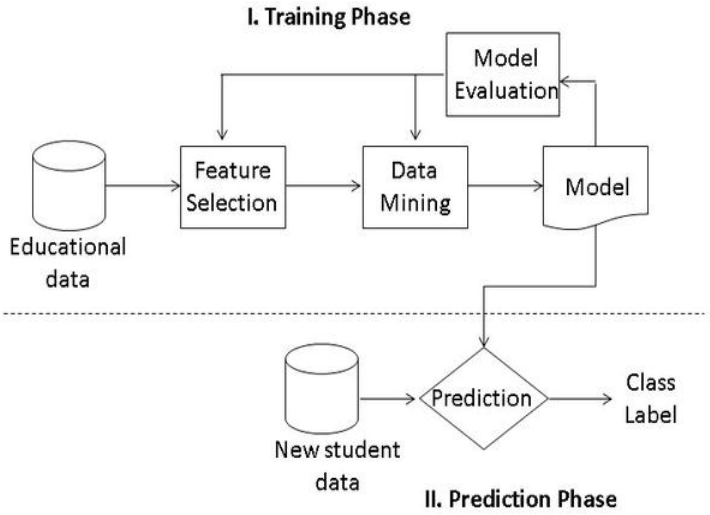

Fig.2: System architecture

The image presents a framework for predicting student academic achievement using data mining techniques, divided into two primary phases: the Training Phase and the Prediction Phase. In the Training Phase, the process begins with a database of Educational Data that includes historical information on students, such as academic scores, attendance records, socio-demographic factors, and other metrics related to student performance. The first step in this phase is Feature Selection, where relevant features or attributes are selected from the dataset. This step helps reduce the dimensionality of the data by identifying only the most impactful variables that influence academic outcomes, making the subsequent data mining steps more efficient and accurate. Following feature selection, the Data Mining step is carried out. Here, various data mining algorithms, such as decision trees, neural networks, and clustering techniques, are applied to the selected features to identify patterns and relationships within the data. This step allows for the extraction of trends that are closely linked with student performance. Based on these findings, a Model is developed. This model is trained to recognize significant patterns and correlations within the dataset, which will later be used to predict future student outcomes. Once

the model is created, it undergoes a Model Evaluation process to assess its accuracy and reliability. This involves using evaluation metrics, such as accuracy, precision, recall, and F1-score, to determine the model's performance. The evaluation step may lead to adjustments in feature selection, data mining techniques, or model parameters to optimize the model's predictive capabilities. After completing the training phase, the Prediction Phase begins. In this phase, a new dataset, referred to as New Student Data, is introduced to the trained model. This data contains similar attributes to those used in the training phase but pertains to students for whom academic predictions are required. The Prediction step involves processing this new data through the trained model, allowing the model to generate predictions based on the patterns it learned in the training phase. The final output of the prediction phase is a Class Label assigned to each student. This label categorizes students according to predicted academic performance or risk level, with classifications that might indicate levels such as "High Risk," "Medium Risk," or "Low Risk" of academic difficulties, or "Excellent," "Good," or "Needs Improvement" based on achievement levels.

4. Implementation

Algorithms

"K-means" pertains to the K-means Algorithm. Each data point should only be associated with one group in the K-means iterative process, which aims to partition the data set into a maximum of ten separate clusters that do not overlap with one another. Identifying clusters in unlabelled data is the job of the K-means clustering method. This has potential applications in both discovering previously unseen groups in massive datasets and verifying business assumptions regarding the types of groups that actually exist.Market researchers might use discriminant analysis to split their data into many categories if necessary, a versatile statistical tool. Assigning objects to one of numerous established categories is what discriminant analysis is all about. One statistical method that employs scores on many The process of using quantitative predictor factors to classify data into separate, non-overlapping groups is called discriminant analysis. One use of discriminant analysis is the identification of patients' risk of stroke, both high and low.Random forest: Data scientists working in fields as diverse as e-commerce, banking, stock trading, and medicine use random forest in their daily work. In order to keep these companies running effectively, it is used to predict things like customer behaviour, medical history, and safety. A network of decision trees is the building block of the random forest approach to categorization. Using bagging and feature randomization, it builds a forest of independent trees with a better collective forecast than any individual tree.One supervised learning classifier is the k-nearest neighbours (or k-NN) approach, which makes use of the idea of proximity to make predictions or classifications about how to group individual data points. It is not a parametric method. Even the most accurate models can't match the accuracy of the KNN algorithm's predictions. Therefore, KNN can be employed in situations where precision is paramount but a human-readable model is not necessary. You can tell how accurate the predictions are by looking at the distance measure.SVM: Classification and regression are two applications of the supervised machine learning method known as the "Support Vector Machine" (SVM). For data points that aren't linearly separable, SVM can classify them by projecting them onto a space with several dimensions. By processing the data in such a way that the hyperplane represents the identified separator between the categories,Machine learning classifications are a kind of estimator that trains many base models or estimators and then aggregates their findings. Coupled vote choices for each estimator output might be aggregation criteria.

The SVM, K-means and KNN are facing limitation with classes balancing and batch normalization issue so that getting more delay. In order to cross over the limitations of above models proposing CNN+LSTM model.

CNN+LSTM: One kind of network design used in deep learning algorithms is convolutional neural networks (CNNs). most often used for pixel data processing and image recognition. While deep learning makes use of a variety of neural network types, convolutional neural networks (CNNs) are best suited for object recognition and identification.The acronym LSTM stands for long short-term memory networks, a tool used in Deep Learning. This particular type of recurrent neural network (RNN) is quite good at sequence prediction and has the ability to learn long-term dependencies.K-Fold is a convolutional neural network (CNN) validation method that uses k subsets and k runs of the holdout technique. The test set is created from one subset, while the training set is made from the other subsets. This method is more reliable than the conventional handout method since it calculates the average error from all k trials.

This study's findings for the field of education are as follows:

1) Tap into the power of the exceptional team to fuel expansion
2) Specifically tailored changes to the curriculum to achieve the aim of skill-based instruction
3) Evaluate more effective methods of instruction to foster student development

CNN+LSTM Algorithm

Input: Sequential data with both spatial and temporal components (e.g., frames of a video, sequence of images, or multivariate time series data)

Output: Predicted class label or forecasted values based on input sequence.

Steps: 1 Data Preparation:

- Collect and preprocess the data: For image sequences, resize and normalize images. For other time-series data, normalize each feature.
- Segment data into sequences: Prepare input sequences, each consisting of multiple steps (e.g., a series of consecutive frames in a video or historical time steps).

Step: 2 Feature Extraction using CNN:

- Input each sequence step to CNN: For each time step in a sequence, pass the data through a CNN model to extract spatial features.
- Convolution and Pooling Layers: Use convolutional layers to capture spatial features (patterns in the image) and pooling layers to reduce dimensionality while retaining essential features.
- Flatten the Output: Flatten the CNN output at each time step into a 1D vector, which represents spatial features for that time step.

Step: 3 Sequence Modeling using LSTM:

- **Combine Sequential CNN Outputs**: Stack the flattened CNN outputs for each time step in the sequence to form a feature sequence.
- **Pass Sequence through LSTM Layers**: Use LSTM layers to capture temporal dependencies across the sequence of CNN-extracted features. LSTM layers will learn sequential relationships and long-term dependencies.

Step: 4 Fully Connected Layers (Dense Layers):

- Add Dense Layers: After the LSTM layers, add fully connected layers to learn higher-level abstractions from the combined spatial-temporal features.
- Output Layer: Use an appropriate activation function in the output layer, such as softmax for classification or linear for regression.

Step :5 Model Training:

- **Define Loss Function**: Choose a suitable loss function, such as categorical cross-entropy for classification tasks or mean squared error for regression tasks.
- **Optimizer**: Select an optimizer (e.g., Adam, RMSprop) and set hyperparameters (learning rate, batch size).
- **Train the Model**: Feed the sequence data into the CNN+LSTM model and train it on the chosen dataset. Adjust the model weights by minimizing the loss function through backpropagation.

Step :6 Model Evaluation:

- Use evaluation metrics such as accuracy for classification or mean squared error for regression to assess the model's performance on a validation/test set.

Prediction:

- Feed new sequences through the trained CNN+LSTM model to generate predictions (e.g., class labels for each sequence or predicted values for time-series forecasting).

The data mining portion of the article begins with clustering student performance using the K-means algorithm from unsupervised learning. Due to the subjective and illogical nature of the school's evaluation results, CNN's category label is derived from the clustering results. Improving the model's optimum forecast accuracy is a major step towards ensuring that schools are being fair and objective when evaluating students. Additionally, students on academic probation might be easily recalled. The clustering number is related to the label value selection range, which is an important consideration when looking at data labels. A well-known problem with the K-means approach is that it uses an arbitrary number for k. By replacing subjective evaluation with quantitative analysis and using the research enhances the procedure & generates more robust clustering results by using an objective statistic to optimise k-value selection. Furthermore, the model's performance is guaranteed by its convincingness, which makes CNN training and prediction outcomes more reliable. Even if the clustering results are obtained after a comprehensive analysis of the present situation and quantitative analysis, the accuracy of the results could be affected because the starting clustering centre is chosen at random. We do not evaluate the proposed statistic against other classifiers, even though it enhances CNN outcomes compared to those without it. There are a number of policy, resource, and technological opportunities for EDM in this era of big data. Education and society benefit greatly from EDM research since it promotes innovation and progress in the field. Because EDM is an interdisciplinary discipline dealing with complex educational challenges, its strength lies in its data sources, data features, research methodology, and application aims. Our mission at EDM is to improve education and the learning process by identifying and resolving research difficulties in the field. We do this by applying a variety of data mining techniques to educational data and by exploiting current data to discover new knowledge. The student dataset is analysed using a hybrid model that incorporates data mining approaches with cutting-edge education data processing tools.

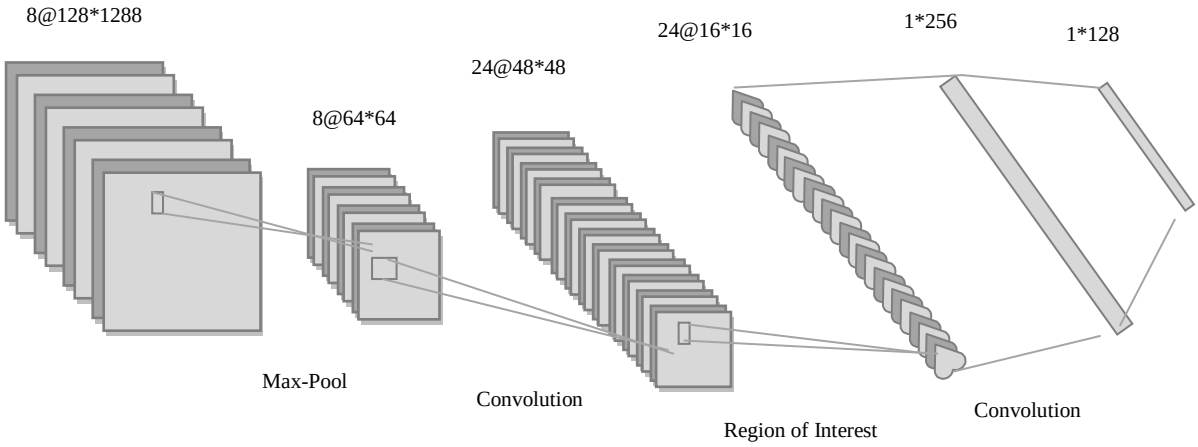

Figure 3: proposed CNN with LSTM architecture

In this paper, a supervisory scheme is implemented with an input layer size of 128x128. The input is divided into 8 sub-sections, each of size 64x64, to reduce the feature range through a max-pooling operation, minimizing the complexity of the feature space. The sub-sections are then processed through convolution and subjected to Region of Interest (ROI) operations. In the second stage of the RCNN, a reverse operation is performed, as shown in Figure 3. The problem is formulated as a segmentation task, defined by equation (1), with each layer's mathematical distribution structured according to the proposed algorithm.

$$I_{seg} = \begin{cases} \text{Normal} & \text{non} - \text{identified Regions} \\ \text{Segmented} & \text{Identified Region} \end{cases} \text{-----(1)}$$

The convolutional layers use a 2D kernel size of 3x3. The network architecture consists of 6 stages and 8 sub-sections, each of size 64x64. Additionally, the max-pooling functionality is further refined by reducing the feature range within the ROI. The forward operation identifies disease-related pixels, while the reverse operation locates and classifies the diseased areas within the test medical images using RCNN.

5. RESULTS AND DISCUSSION

In this section a brief discussion of Forecasting Student Academic Achievement on Educational Data with Mining and deep learning Techniques were explained.

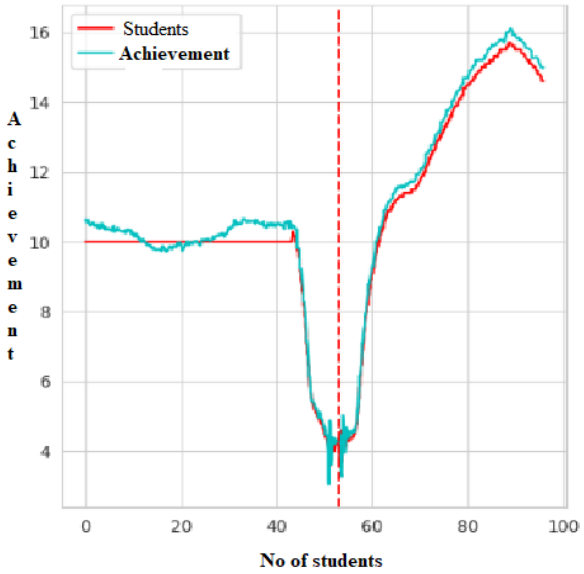

Figure 3: students vs achievements training

The above figure 3 clearly explains about students count vs achievements score training analysis, in this proposed model efficiently attains more accuracy and sensitivity.

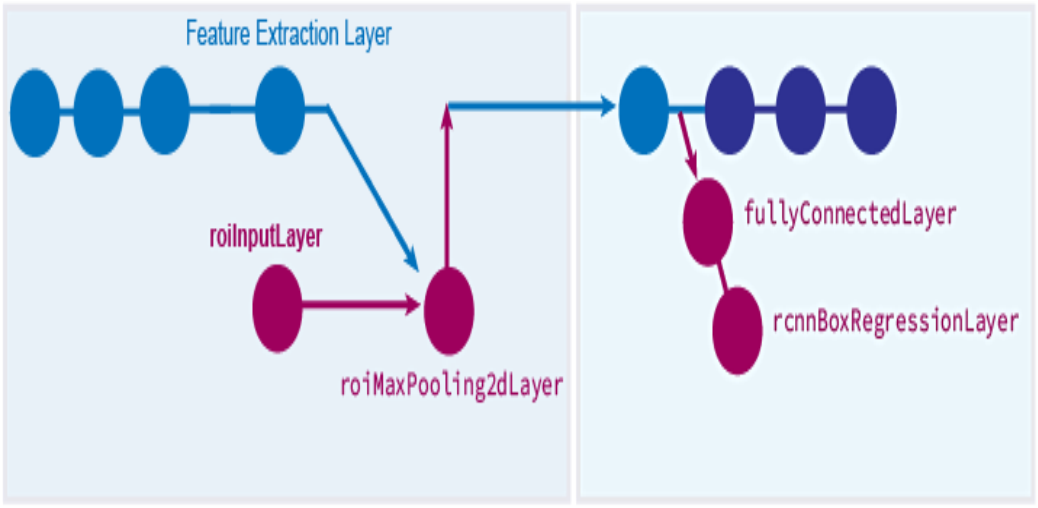

Figure 4: Layered wise operation

The notation used includes several parameters: X and Y represent the spatial coordinates, while L, W, and H denote the length, width, and height, respectively. The variable n stands for the feature count, and ccc indicates the number of channels. The filter count is represented by f, and S denotes the stride length. Additionally, BS is used to signify the batch size, and Reg represents regression. During pooling operations, image patches are verified to check whether they correspond to the Region of Interest (ROI), as shown in Figure 4.

To assess the performance of the proposed method in classifying cardiovascular disease information, several key metrics are used, including sensitivity, precision, efficiency, ROC curve, recall, accuracy, and F1 score. Each of these metrics plays an essential role in evaluating the classifier's effectiveness.

$$\text{Sensitivity} = \text{TNR} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

$$\text{Specificity} = \text{TPR} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \quad (3)$$

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (6)$$

$$\text{F}_1 = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}} \quad (7)$$

The performance measures used demonstrate the robustness and strength of the application. Metrics such as sensitivity (Eq. 2), specificity (Eq. 3), accuracy (Eq. 4), recall (Eq. 5), precision (Eq. 6), and F1 score (Eq. 7) are evaluated using the functionality of the confusion matrix. The true positive rate, false positive rate, true negative rate, and false negative rate are key factors in determining the reliability of these metrics. According to the RCNN model, these performance measures have shown improvement over previous models, highlighting the enhanced effectiveness of the approach [26].

Table 1: Measures analysis comparison

Techniques	**TP+TN**	**TP+TN+FP+FN**	**AUC (%)**
ARF [27]	19890	20112	90.76
SVM [28]	12,780	20134	68.27
DT [29]	13,633	20235	67.787
KNN [30]	14,489	20134	69.216
CNN+LSTM (proposed)	19,607	20137	99.451

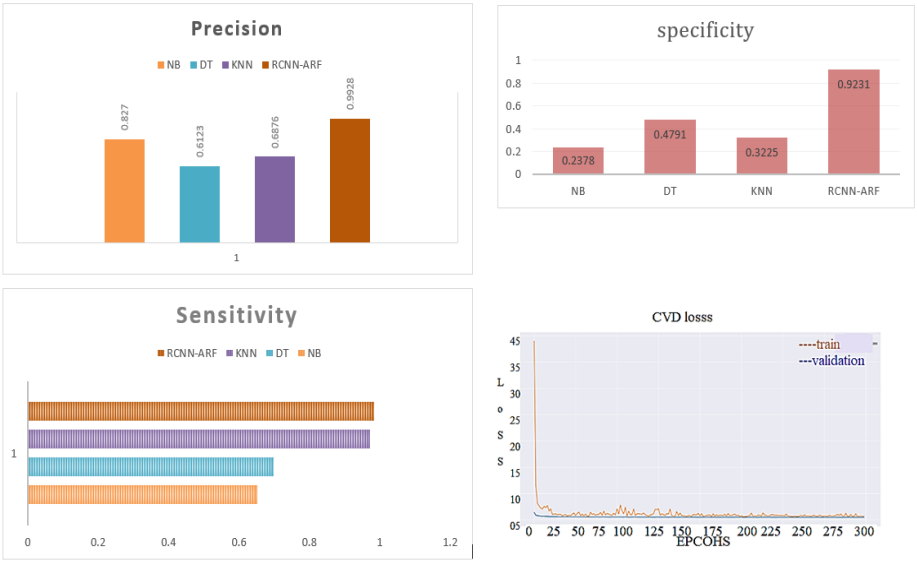

Figure 5: All performance measures using confusion matrix

Figure 5 provides a clear illustration of the accuracy analysis for the proposed model, where the deep learning-based CNN+LSTM approach demonstrates a high level of performance. While earlier models achieved an accuracy level around 70%, the implemented CNN+LSTM model has improved classification accuracy to exceed 80%. This significant enhancement highlights the robustness and precision of the proposed method over previous approaches.

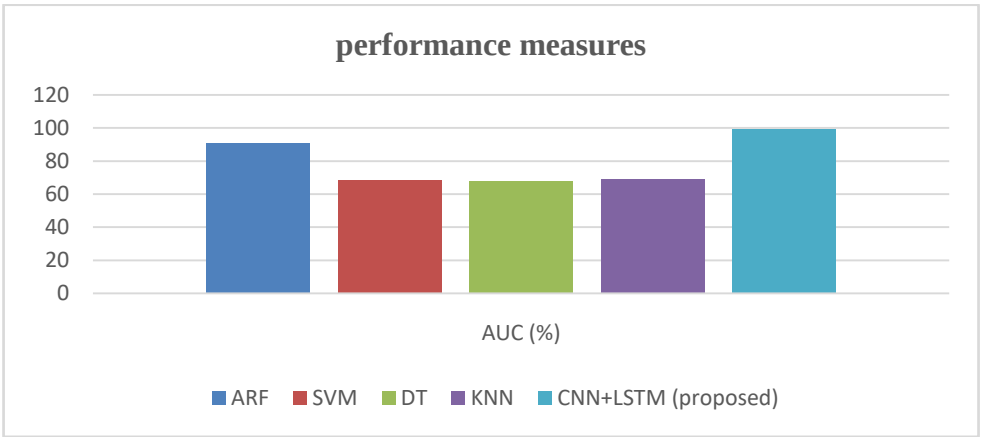

Figure 5: comparisons of models

To evaluate and compare the effectiveness of various models, performance measures such as sensitivity, specificity, accuracy, recall, precision, and F1 score are analyzed.

6. CONCLUSIONS

This study demonstrates the effective application of advanced data mining techniques and intelligent technologies to improve the accuracy and reliability of academic performance forecasting. Conventional processing methods fall short in handling the vast and complex data generated in educational settings, making intelligent techniques essential. By combining grouping theories, discriminant analysis, and convolutional neural networks (CNN), the proposed approach effectively minimizes inaccuracies in student assessments and enables predictive analysis of future academic performance.The study introduces a novel statistic to optimize the determination of clustering numbers in the K-means algorithm, enhancing the model's clustering efficacy as verified through discriminant analysis. The CNN model, trained and tested with labeled data, provides a robust tool for forecasting academic outcomes. Cross-validation results show an impressive accuracy of 98.34% and a recall of 97.85%, verifying the model’s predictive strength. These findings confirm that the proposed model not only addresses key challenges in educational data processing but also offers a reliable approach to student performance prediction, with significant potential for application in educational planning and personalized learning strategies.

7. FUTURE SCOPE

The future might hold better results if integration-based technologies or association models were to be combined. There are a variety of industries that can benefit from EDM, including medical data processing and sports data processing. Possible topics for future research include tracking students' emotional well-being, analysing student performance in online courses, developing strategies to promote self-discipline, and utilising educational data mining technologies to find ways to improve student conduct in the classroom. We are encouraged to continue our research because data mining technologies are important for forecasting academic achievement and boosting learning capacity.

REFERENCES

1. H. B. Antonio, H. F. Boris, T. David, and N. C. Borja, ‘‘A systematic review of deep learning approaches to educational data mining,’’ Complexity, vol. 2019, May 2019, Art. no. 1306039.
2. M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, ‘‘Implementing AutoML in educational data mining for prediction tasks,’’ Appl. Sci., vol. 10, no. 1, pp. 90–117, Jan. 2020.
3. S. Kausar, X. Huahu, I. Hussain, W. Zhu, and M. Zahid, ‘‘Integration of data mining clustering approach in the personalized E-learning system,’’ IEEE Access, vol. 6, pp. 72724–72734, 2018.
4. D. Buenaæo-Fernandez, W. Villegas, and S. LujÆn-Mora, ‘‘The use of tools of data mining to decision making in engineering education—A systematic mapping study,’’ Comput. Appl. Eng. Educ., vol. 27, no. 3, pp. 744–758, May 2019.
5. C. Angeli, S. K. Howard, J. Ma, J. Yang, and P. A. Kirschner, ‘‘Data mining in educational technology classroom research: Can it make a contribution?’’ Comput. Educ., vol. 113, pp. 226–242, Oct. 2017.
6. B. A. Javier, F. B. Claire, and S. Isaac, ‘‘Data mining in foreign language learning,’’ WIREs Data Mining Knowl. Discov., vol. 10, no. 1, Jan./Feb. 2020, Art. no. e1287.
7. C. Romero and S. Ventura, ‘‘Data mining in education,’’ Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, vol. 3, no. 1, pp. 12–27, 2013.
8. C. Romero and S. Ventura, ‘‘Educational data mining and learning analytics: An updated survey,’’ WIREs Data Mining Knowl. Discov., vol. 10, no. 1, May 2020, Art. no. e1355.
9. S. Wang, ‘‘Smart data mining algorithm for intelligent education,’’ J. Intell. Fuzzy Syst., vol. 37, no. 1, pp. 9–16, Jul. 2019.
10. M. J. James, S. H. Ganesh, M. L. P. Felciah, and A. K. Shafreenbanu, ‘‘Discovering students’ academic performance based on GPA using K-means clustering algorithm,’’ in Proc. World Congr. Comput. Commun. Technol., Trichirappalli, India, 2014, pp. 200–202.
11. A. Ani, L. Nicholas, and S. B. Ryan, ‘‘Enhancing the clustering of student performance using the variation in confidence,’’ in Proc. Int. Conf. Intell. Tutoring Syst. Cham, Switzerland: Springer, 2018, pp. 274–279.
12. R. G. Moises, D. P. P. R. Maria, and O. Francisco, ‘‘Massive LMS log data analysis for the early prediction of course-agnostic student performance,’’ Comput. Educ., vol. 163, Apr. 2020, Art. no. 104108.
13. J. N. Walsh and A. Rísquez, ‘‘Using cluster analysis to explore the engagement with a flipped classroom of native and non-native Englishspeaking management students,’’ Int. J. Manage. Educ., vol. 18, no. 2, Jul. 2020, Art. no. 100381.
14. V. G. Karthikeyan, P. Thangaraj, and S. Karthik, ‘‘Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation,’’ Soft Comput., vol. 24, no. 24, pp. 18477–18487, Dec. 2020.
15. L. M. Crivei, G. Czibula, G. Ciubotariu, and M. Dindelegan, ‘‘Unsupervised learning based mining of academic data sets for students’ performance analysis,’’ in Proc. IEEE 14th Int. Symp. Appl. Comput. Intell. Informat. (SACI), Timisoara, Romania, May 2020, pp. 11–16.
16. Ramaiah, V. S., Singh, B., Raju, A. R., Reddy, G. N., Saikumar, K., & Ratnayake, D. (2021, March). Teaching and Learning based 5G cognitive radio application for future application. In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 31-36). IEEE.
17. Mohammad, M. N., Kumari, C. U., Murthy, A. S. D., Jagan, B. O. L., & Saikumar, K. (2021). Implementation of online and offline product selection system using FCNN deep learning: Product analysis. Materials Today: Proceedings, 45, 2171-2178.
18. Padmini, G. R., Rajesh, O., Raghu, K., Sree, N. M., & Apurva, C. (2021, March). Design and Analysis of 8-bit ripple Carry Adder using nine Transistor Full Adder. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1982-1987). IEEE.
19. Dr. k. Raju, A. Sampath Dakshina Murthy, Dr. B. Chinna Rao, Sindhura Bhargavi, G. Jagga Rao, K. Madhu, K. Saikumar. A Robust And Accurate Video Watermarking System Based On SVD Hybridation For Performance Assessment International Journal of Engineering Trends and Technology, 68(7),19-24.

20. Saba, S. S., Sreelakshmi, D., Kumar, P. S., Kumar, K. S., & Saba, S. R. (2020). Logistic regression machine learning algorithm on MRI brain image for fast and accurate diagnosis. International Journal of Scientific and Technology Research, 9(3), 7076-7081.
21. Saikumar, K. (2020). RajeshV. Coronary blockage of artery for Heart diagnosis with DT Artificial Intelligence Algorithm. Int J Res Pharma Sci, 11(1), 471-479.
22. Saikumar, K., Rajesh, V. (2020). A novel implementation heart diagnosis system based on random forest machine learning technique International Journal of Pharmaceutical Research 12, pp. 3904-3916.
23. Raju K., Chinna Rao B., Saikumar K., Lakshman Pratap N. (2022) An Optimal Hybrid Solution to Local and Global Facial Recognition Through Machine Learning. In: Kumar P., Obaid A.J., Cengiz K., Khanna A., Balas V.E. (eds) A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems. Intelligent Systems Reference Library, vol 210. Springer, Cham. https://doi.org/10.1007/978-3-030-76653-5_11
24. Sankara Babu B., Nalajala S., Sarada K., Muniraju Naidu V., Yamsani N., Saikumar K. (2022) Machine Learning Based Online Handwritten Telugu Letters Recognition for Different Domains. In: Kumar P., Obaid A.J., Cengiz K., Khanna A., Balas V.E. (eds) A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems. Intelligent Systems Reference Library, vol 210. Springer, Cham. https://doi.org/10.1007/978-3-030-76653-5_12
25. Kiran Kumar M., Kranthi Kumar S., Kalpana E., Srikanth D., Saikumar K. (2022) A Novel Implementation of Linux Based Android Platform for Client and Server. In: Kumar P., Obaid A.J., Cengiz K., Khanna A., Balas V.E. (eds) A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems. Intelligent Systems Reference Library, vol 210. Springer, Cham. https://doi.org/10.1007/978-3-030-76653-5_8 .
26. K. Kishore Raju, Ch.S.V.V.S.N. Murty, Suresh Kumar Kanaparthi, Amdewar Godavari, Kayam Saikumar, "Multi-Dimensional Machine Intelligence Technique on High Computational Data for Bigdata Analytics," SSRG International Journal of Electrical and Electronics Engineering, vol. 11, no. 6, pp. 91-100, 2024. Crossref, https://doi.org/10.14445/23488379/IJEEE-V11I6P110
27. Yamsani, N., Sarada, K., Ahmed, M. A., & Saikumar, K. (2024, March). Estimate and prevention of malicious URL using logistic regression ML techniques. In AIP Conference Proceedings (Vol. 2919, No. 1). AIP Publishing.
28. Revathi, R., Kumar, K. S., Hamid, H. A., & Ahammad, H. S. (2024, March). A real time homogeneous wireless sensor network functioning using semi-tier clustering protocol for energy efficiency routing. In AIP Conference Proceedings (Vol. 2919, No. 1). AIP Publishing.
29. Bommagani, N. J., Venkataramana, A., Vemulapalli, R., Singasani, T. R., Pani, A. K., Challageri, M. B., & Kayam, S. (2024). Artificial Butterfly Optimizer Based Two-Layer Convolutional Neural Network with Polarized Attention Mechanism for Human Activity Recognition. Mathematical Modelling of Engineering Problems, 11(3).
30. Kumar, N. K., Sarada, C., Mallika, S. S., Subhashini, P., & Zearah, S. A. (2024, April). Human Blood Cell (Haematologicalimages) Recognition by Applying Deep Convolution Methods. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.