I.J. Modern Education and Computer Science, **2025, 1, 59-71**
Published Online on February 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2025.01.05

Canberra Match Normalization-Enhanced Decision Stump Classifier for Predicting Academic Performance in the Context of Smartphone Addiction

R. Ruth Belina*
Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli - 620002, Tamil Nadu, India
Email: r.ruthbelina@gmail.com
ORCID iD: https://orcid.org//0000-0002-7444-6326
*Corresponding Author

T. Lucia Agnes Beena
Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli - 620002, Tamil Nadu, India
ORCID iD: https://orcid.org//0000-0001-5598-3962

Charles Savarimuthu
IT Department, University of Technology and Applied Sciences - Al Mussanah, Muladdah, Mussanah, P.O. Box: 191 | Postal Code: 314, South Batina, Sultanate of Oman
ORCID iD: https://orcid.org//0000-0003-3680-8135

Received: 11 December, 2023; Revised: 16 February, 2024; Accepted: 07 October, 2024; Published: 08 February, 2025

Abstract: Student academic performance (SAP) prediction is a key issue in education data analysis. Also, the assessment of students' performance is used to enhance the efficiency of educational institutions. With the development in educational institutions and modern technology, focusing on the academic performance prediction of the student based on access to the smartphone is the need of the hour. To improve the accuracy of student academic performance prediction, the Canberra Match Normalization-based Generalized Canonical Correlative Decision Stump Classifier (CMN-GCCDSC) is introduced. Initially, student data are collected from the dataset. After the data collection process, the proposed CMN-GCCDSC technique is applied in two phases namely data preprocessing and classification respectively. In the first phase, data preprocessing is carried out to eliminate duplicate data using the Canberra Match Data Normalization technique to minimize space and time consumption. In the second phase, data classification is performed with preprocessed output to classify student academic performance using a generalized canonical correlative decision stump classifier based on Smartphone addiction prediction. The generalized canonical correlation analysis is used for decision-making. Based on analysis, student academic performance is classified and results are obtained. An experimental assessment of the proposed CMN-GCCDSC technique and existing methods is carried out with metrics such as accuracy, sensitivity, specificity, space complexity, and time complexity. The CMN-GCCDSC technique is an effective solution that addresses the limitations of Genetic Algorithm (GA)-based decision tree classifiers. By combining the Decision Stump Classifier (DSC) approach with Generalized Canonical Correlation (GCC), the most important feature to consider for academic prediction among students can be selected, ultimately reducing the dimensionality of the dataset, and improving classifier performance. With higher accuracy rates achieved, this technique can help identify at-risk students early and discover hidden trends and patterns in student performance, leading to improved academic outcomes with additional support from institutions and faculties.

Index Terms: Student Academic Performance Prediction, Data Preprocessing, Canberra Match Data Normalization Technique, Generalized Canonical Correlative Decision Stump Classifier, Decision Rule

1. Introduction

Student academic performance has become a prevalent issue, as an increasing number of individuals depend heavily on their smartphone usage and other activities. In the educational domain, the rapid increase of Smartphones and their extensive usage among students have raised concerns about their impact on academic performance. Digital learning has numerous positive outcomes, including interactive and flexible learning, convenience, and personalized learning. It also has environmental benefits, such as reducing paper waste and the number of trees cut down. However, it can also have negative impacts, such as technical barriers, distraction, isolation, and dependence on technology. Digital fasting can help reduce the negative impact of excessive screen time on our mental and physical health.

Smartphones have been linked to a decline in academic performance, increased stress, and poor sleep. The mere presence of a smartphone can be a distraction that affects an individual's fluid intelligence and working memory capacity, leading to poor attention. Educational Data Mining is a recent technology that has emerged to predict and analyze the impact of smartphones on learning outcomes. Now educational researchers recognize the potential to improve learning by leveraging the data collected from students' smartphone usage.

1.1 Aim and Scope of the CMN-GCCDSC Method

Smartphones can be an invaluable tool for students, providing more than just the ability to text and call. With instant access to the internet, students can use their phones to research topics, find resources for assignments, and access educational websites. Additionally, mobile phones offer opportunities to participate in online courses, webinars, and educational apps, giving students access to a more flexible and convenient learning experience. Moreover, mobile phones can offer a source of relaxation and entertainment during study breaks, which can help reduce stress and improve overall well-being. However, students must use their mobile phones responsibly to avoid distractions and negative impacts on academic performance. It's recommended to establish boundaries for smartphone usage during study sessions and classes to ensure maximum focus and productivity.

Our study aims to predict smartphone addiction among students by analyzing data collected through various classifiers and techniques. This research will not only enhance the accuracy of predictions of smartphone addiction but also provide valuable insights into the impact of smartphone usage on academic results. The results can be used by faculty and student counselors to raise awareness regarding the issue. Moreover, the predictions can help educational institutions to identify students who are at risk due to excessive smartphone usage, which will help improve their attention span during class and learning speed.

This paper is structured as follows: Section 2 reviews the related works on student performance prediction. Section 3 provides a concise explanation of the proposed CMN-GCCDSC technique. Section 4 describes the experimental setup and performance results of the CMN-GCCDSC technique. Section 5 summarizes the key findings in the experiential results.

The novelty of the proposed CMN-GCCDSC technique

The main novelty of this proposed CMN-GCCDSC technique is summarized as follows.

- Proposed CMN-GCCDSC is introduced to enhance the accuracy of student academic performance prediction.
- The Canberra Match Data Normalization Process, a novel approach is executed to pre-process the data for removing the repeated data from the database. Every data is checked with another data to find repeated one for minimizing the dimensionality. This novel technique was introduced to reduce space and time consumption.
- Another innovative technique, the generalized canonical correlative decision stump classifier is implemented in preprocessed output to classify student academic performance based on Smartphone addictive prediction.
- Finally, performance analysis is conducted for the CMN-GCCDSC technique through experiments to verify the efficiency of the method in terms of accuracy, sensitivity, specificity, space complexity, and time complexity based on the number of student academic performances.

1.2 Research Gap:

Improving data collection accuracy, sensitivity, and specificity while reducing time and space complexity is the main goal of our research. Although Genetic algorithms have been used in other studies, they are not a scalable solution for complex data sets. Stacked models, on the other hand, require a significant amount of computing power and can be time-consuming and expensive. Simple random sampling is another method used in different studies but it can be more time-consuming and expensive, and the sample selected may not represent the population. Artificial Neural Network is computationally intensive and resource-consuming, requiring large amounts of labeled training data. Deep Neural Networks (DNN) are also not without weaknesses, as they require a high amount of data to train them, and input data must provide greater variation to prevent "overfitting." Unless a project has access to significant computing power, DNN often fails to provide superior output over conventional methods. Our current research tackles these concerns by using the CMN-GCCDSC model, which is a more time-efficient and cost-effective solution.

2. Related Work

A Genetic Algorithm (GA)-based decision tree classifier was developed [1] to predict students' academic performance with higher accuracy and minimal errors. But, the time and space complexity was a challenging issue. A stacking algorithm was introduced [2]predicting the academic performance of middle-and high-school students. But the algorithm failed to enhance the sensitivity performance. A mediation and moderation variable analysis was conducted [3] for Smartphone addiction prediction on students' academic performance. However, it failed to enhance prediction accuracy. Pearson correlation analysis and one-way multivariate ANOVA were applied [4]to differentiate the relationship between smartphone behavior and academic performance. But more robust models were not employed. A simple random sampling method was employed [5]to explore the correlation between smartphone addiction and academic performance. However, it was not efficient for predicting the academic performance. A deep neural network was designed [6]for predicting the students' academic performance. The network was minimized Mean Absolute Error and Root Mean Squared Error, but the time complexity was increased. An artificial neural network was introduced [7]for student predicting academic performance in higher education. However, it provides the inaccurate predictions. An attention-based Bidirectional Long Short-Term Memory (BiLSTM) network was designed [8] to efficiently predict student performance. The sensitivity of the prediction was not improved. A deep neural network model was designed [9]to predict students' performance at an early stage. But the accuracy of the model was not improved. The hybrid regression model was developed [10]to optimize the prediction accuracy of student academic performance. However, the designed model failed to set optimal thresholds for classifying the data.

An Augmented Education (Augment ED) model was designed [11] to accurately predict student academic performance with a high level of accuracy. However, the designed model required more time for prediction. A t-Distributed Stochastic Neighbor embedding (t-SNE) technique was developed [12] to enhance the prediction of students' academic performance at an early stage. However, the deep learning architecture was not improved performance prediction. A collaborative learning approach was designed [13] to analyze and track students' academic performance results. However, the approach provided inefficient results when dealing with large datasets. A new Tri-Branch CNN architecture was developed [14]to accurately detect students' academic performance. However, it failed to collect more relevant data with students' behavioral records. The convolutional neural network was introduced [15]for training and testing data analysis for academic performance prediction. However, the space complexity was not reduced.

The deep learning model was introduced [16] early prediction of student performance, enabling timely intervention by university to implement corrective strategies for students support and counselling. But the time complexity was not minimized. A Deep Neural Network (DNN) framework was designed [17] for binary classification with evaluated by using several activation functions and optimizer. But it failed to improve the prediction accuracy. The Internet usage behaviors and academic performance was designed[18] to predict undergraduate's academic performance from the usage data by machine learning. However, along with improve the download volume Internet connection duration and academic performance decreases. A novel prediction algorithm for evaluating student's performance was introduced[19] to develop based on both classification and clustering techniques. It improves the process of learning and minimum the failure rates of academics. The Students' Academic Performance Enhancement (SAPE) was performed[20] for predicting students' academic performance to predict students with learning difficulties by their homework submission behaviors. But the furcating accuracy was not enhanced by designed SAPE method.

A novel model for predicting university student performance based on multi-feature fusion and attention mechanisms was introduced [21] to focuses on analyzing historical academic grades from multiple dimensions among university students to extract features. However, difficult to collect many relevant factors other than past academic performance. A web-based system for predicting academic performance and identifying students at risk of failure through academic and demographic factors is developed [22]. But, it was contributed to determine the new factor that has not been studied in previous studies, which is living place (city). An Ensemble Meta-Based Tree Model was performed [23]to build the prediction model by different families of Machine Learning Techniques on selected dataset for consideration. A machine learning technique were introduced [24]to predict the final student grades in 1st semester courses by improving the performance of predictive accuracy. However, it was different challenges in handling imbalanced datasets for enhancing the performance of predicting student grades. An empirical performance and efficiency of ensemble classification methods were designed [25] for successful predictions. The student academic performance prediction utilizes machine and deep learning models, but ensemble models are not very well investigated.

3. Proposal Methodology

Smartphone's have become a well-turned medium for sharing interests in the course of social as well as academic connections. Usage of the Smartphone in higher education organizations declares that students communicate with each other via their cell phones to transfer notes, lectures, projects and etc. Apart from that, students are also involved in social interactions and game-related usage. The high level of addiction to social interactions and games on Smartphone harmfully influences the student's academic. Therefore, a novel technique CMN-GCCDSC is introduced with two phases for enhancing the student academic performance prediction based on the Smartphone's addiction.

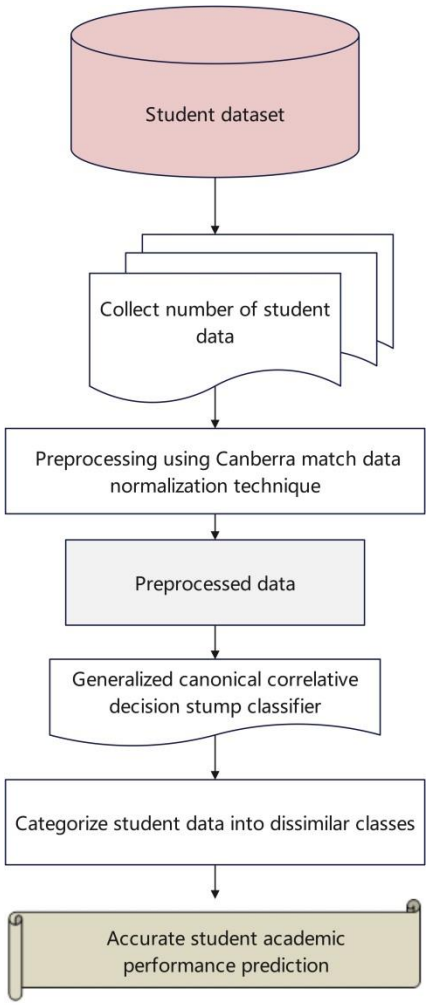

Fig. 1. Flow chart of the proposed CMN- GCCDSC technique

Figure 1 exhibits the flow chart of the proposed CMN-GCCDSC technique for accurate student academic performance prediction based on Smartphone addiction. The real time student dataset is considered as input to the proposed technique. The numbers of student data '$d_1, d_2, d_3 \ldots\ldots d_m$' are collected from this dataset. The proposed CMN-GCCDSC technique uses the Canberra match data normalization technique for data preprocessing. After the data preprocessing, generalized canonical correlative decision stump classifier is employed in proposed CMN-GCCDSC technique for analyzing the preprocessed data. Based on the analysis, student data is categorized into different classes.

3.1 Dataset description

The student dataset was collected through a survey involving 1,115 students from Arts & Science Colleges. The dataset consists of four parts such as Sociodemographic Details, Psychological Factors, Academic Performance Factors, and Social Factors. In this work, a total of38 attributes are used in the form of questions based on the smart phone usage of the students. Also, 16 smartphone addiction attributes and other relevant attributes are utilized. By using smartphone addiction attributes and other attributes, students' academic performance are classified.

3.2 Canberra match data normalization technique-based data preprocessing

Data preprocessing is the fundamental step used for cleaning and transforming the raw dataset into a structured format. Generally, the dataset contains missing values, duplicate values and unusable format which are not directly used for machine learning standards to perform classification tasks. The duplicate value in the dataset affects the accuracy of the classification task. Therefore, data preprocessing is a required task for cleaning the data and making it suitable format for a machine learning model to increase the accuracy and efficiency of a classification process. A novel Canberra Match Data Normalization Process is performed to pre-process the data for removing the repeated data from database. Every data is checked with another data to find repeated one for minimizing the dimensionality. This novel technique reduced the space consumption and time consumption. The Data preprocessing using Canberra match data normalization is a distance measure used to calculate the distance between two points based on the differences between their corresponding components. It is useful when comparing data that is not necessarily normalized or standardized.

The proposed CMN-GCCDSC technique uses the Canberra match data normalization technique for data preprocessing to remove duplicate data. The Canberra match data normalization technique is a numerical measure of the distance between the data in a given dataset. To make them suitable for a machine learning model, the data is converted into an appropriate format before applying the techniques.

To start with the raw input dataset 'D' and data are formulated as in eq (1).

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \dots & \vdots \\ d_{m1} & d_{m2} & \dots & d_{mn} \end{bmatrix} \quad (1)$$

From the input dataset formulation as given in eq (1), the dataset 'D' separated by 'm' rows and 'n'columns. In this format, column 'n' represents the number of features or attributes and the 'm' rows denote an overall sample instances or records or student data. From eq (1), the student data is represented as$d_1, d_2, d_3 \dots . . d_m$.

With the above-said data format, data preprocessing is performed. To acquire the structured dataset, the Canberra Match data normalization process is performed to identify the duplicate student data by measuring the linear relationship between the student's performance data. The relationship is measured between the data is measured as given below,

$$CS = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|d_i - d_j|}{|d_i| + |d_j|} \quad (2)$$

Where, 'CS' indicates the Canberra similarity between the data as given in eq (2), 'd_i' denotes a i^{th} student data, 'd_j' denotes the j^{th} student data i.e. neighboring data, $|d_i - d_j|$ denotes a distance between the data,$|d_i|$ and$|d_j|$ denotes a cardinality of a set which defined as the number of elements in a set. The similarity value lies between the values '0' and '1'.

$$CS = \begin{cases} (0 \; to \; 0.5), & duplicate \;\; data \\ (0.5 \; to \; 1), & normal \;\; data \end{cases} \quad (3)$$

Where, CS provides the output as a duplicate data if the similarity values from 0 to 0.5 in eq (3). If the similarity values from 0.5 to 1, then it is said to be normal data. Depending on similarity value, the duplicate student data values are identified from the dataset and it removed. The Canberra match data normalization-based data preprocessing algorithm 1 is given in Table 1.

3.3 Generalized canonical correlative decision stump classifier

After the preprocessing, classification phase is performed by Generalized Canonical Correlative Decision Stump Classifier (GCCDSC) for student academic performance prediction based on Smartphone addiction. An innovative method, generalized canonical correlative decision stump classifier is implemented on preprocessed output to classify student academic performance based on Smartphone addiction prediction. A decision stump classifier is a machine-learning technique that uses the Generalized Canonical Correlation (GCC) to measure the relationship between a dependent variable and the independent variable. The dependent data is considered as student academic performance. The independent data is considered as mobile addiction. The student academic performance often called the 'outcome' or 'response' variable or labels such as Marginal, Below Average, Average, Above Average, Good, and Excellent. The technique includes more mobile addiction factors often called 'features' or attributes such as Smartphone addiction attributes and other attributes in the dataset. The academic performance(d_R) is derived based on the summed score of the Smartphone addiction attributes and other attribute values. The mobile addiction is used to evaluate the final performance of an academic outcome. By using Smartphone addiction attributes, mobile addiction (d_T) value in academic performance of classification is less than 50%, 50-60%, 60-70%, 70-80%, 80-90%, and above 90%.Smartphone addiction attributes are employed to calculate the student academic performance. According to the calculation, the student academic performance classified based on different levelssuch as Marginal, Below Average, Average, Above Average, Good and Excellent.

Figure 2 illustrates the flow graph of the GCCDSC for student academic performance prediction, which aims to achieve higher accuracy. The classifier begins by receiving the preprocessed data. Decision stump is one of the most straightforward classification algorithms. A decision stump is a simple decision tree model with consists of one decision node i.e. root node and two leaf nodes. It is a basic form of decision tree that is used for classification tasks. It works by generating a decision tree with only one level, named a "stump". Every stump includes a single feature and a threshold value which divides the data into two groups depending on whether they meet the condition or not. The decision stump classifier utilizes canonical correlation analysis. This correlation analysis is a statistical technique employed to measure the relationships between the student academic performance and mobile phone addiction. By applying this analysis, the

student data is classified into respective class's academic performance. Let us consider the number of student data $d_R = d_1, d_2, d_3 \ldots\ldots d_m$.

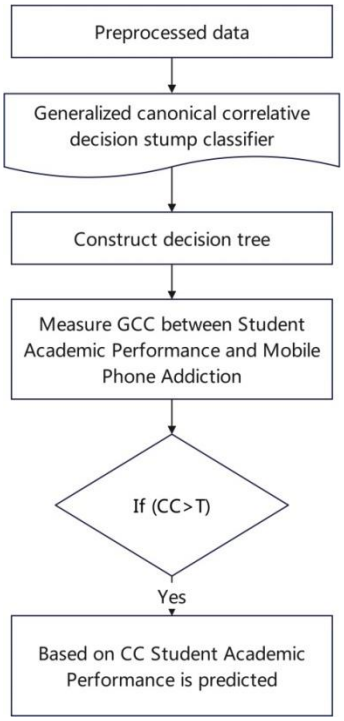

Fig. 2. Flow Graph of GCCDSC

Table 1. Canberra match data normalization-based data preprocessing algorithm

Canberra match data normalization-based data preprocessing
Input: Dataset 'D', Number of attributes or features '$A_j = A_1, A_2, A_3, \ldots\ldots A_n$', data $d_1, d_2, d_3 \ldots\ldots d_m$
Output: Preprocessed data
Begin **Step 1**: Collect number of data and attributes **Step 2:** Formulate input data with row and column using $D = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ d_{m1} & d_{m2} & \ldots & d_{mn} \end{bmatrix}$ **Step 3: For each data** d_i **Step 4: For each data** d_j **Step5**: Measure the Canberra similarity $CS = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\lvert d_i - d_j \rvert}{\lvert d_i \rvert + \lvert d_j \rvert}$ **Step 6:** $\boldsymbol{if}$ $(CS \geq 0 \;\&\&\; CS < 0.5)$ **then** **Step 7:** data is said to be a duplicate and removed **Step 8: else**$\boldsymbol{if}$ $(CS \geq 0.5 \;\&\&\; CS \leq 1.0)$ **then** **Step 9:** data is said to be a normal **Step 10: End if** **Step 11: End for** **Step 12: End for** **End**

The algorithm step of Canberra match data normalization-based data preprocessing is described in Table 1 where the entire Dataset collected is considered as input and using the Canberra match data normalization-based preprocessing, redundant data is removed. The correlation is measured as follows,

$$CC = e^{\left[-\frac{|d_R - d_T|^2}{2*D^2}\right]} \tag{4}$$

$$D = \sqrt{\frac{|d_R - d_T|^2}{n}} \tag{5}$$

Where CC denotes a canonical correlation coefficient between academic performance 'd_R' and mobile addiction 'd_T', 'e' denotes an exponential function in eq (4), 'D' represents the deviation, 'n' denotes a number of data in eq (5). The correlation coefficient 'CC' provides results ranging from 0 to 1.

After measuring the correlation, the decision rule is formed at the root node for classification of the student academic performance based on Smartphone addiction and other attributes. The decision rule is formed as follows,

$$Y = \begin{cases} CC > T, & Student\ academic\ performance \\ & is\ correctly\ classified \\ CC < T, & Student\ academic\ performance \\ & is\ incorrectly\ classified \end{cases} \quad (6)$$

Where, Y denotes an output of the decision tree classifier, 'T' indicates a threshold for correlation coefficient 'CC' results in eq (6). If the coefficient $'CC'$ exceeds the threshold, student academic performance is correctly classified. Finally, the classification results are obtained at the leaf node of the decision tree. In this way, student academic performance prediction is obtained with higher accuracy. The GCCDSC algorithm 2 is shown in Table 2.

Table 2. Generalized canonical correlative decision stump classifier algorithm

Generalized canonical correlative decision stump classifier
Input: Number of student data $d_R = d_1, d_2, \ldots, d_m$., mobile addiction$d_T$ **Output:** Increase the academic performance of classification accuracy
Begin **Step 1: For** each student data d_R **Step 2: For each** mobile addiction 'd_T' **Step 3**: Construct decision tree **Step 4:**Root node measure canonical correlation$CC = e^{\left[-\frac{\lvert d_R - d_T\rvert^2}{2*D^2}\right]}$ **Step 5:** Measure deviation $D = \sqrt{\frac{\lvert d_R - d_T\rvert^2}{n}}$ **Step 6**: $\boldsymbol{If}(CC > T)$**then** **Step 7**: Student academic performance is correctly Classified **Step 8:else** **Step 9**:Student academic performance is incorrectly Classified **Step 10**: **End if** **Step 11: End for** **Step 12: End for** **End**

Table 2 illustrates the algorithmic description of generalized canonical correlative decision stump classifier where the number of student data collected for the mobile addiction will be considered as an input to arrive at accuracy of classification.

4. Experimental Evaluation and Results Analysis

Experimental assessment of proposed CMN-GCCDSC is implemented by Python programming language. For the experimental consideration, the number of input data is taken in the range of 200, to 1115 in steps of 200. The performances of proposed and existing methods are evaluated using different metrics such as sensitivity, specificity, space complexity and time complexity with respect to student data.

4.1 Space Complexity

Space complexity is measured as the amount of memory space consumed for classifying the student academic performance prediction. The formula for calculating specificity is given in eq (7).

$$SC = Number\ of\ student\ data * Space\ consumed\ by\ one\ data \quad (7)$$

Where, 'SC' indicates a space complexity measured in terms of megabytes (MB).

Table 3. Performance results of space Complexity

Number of Student Data (Number)	**Space Complexity (MB)**	
	GA based decision tree classifier	**CMN-GCCDSC**
200	24	18
400	25	20
600	30	27
800	37	31
1000	39	34
1115	47	39

Table 3provides the performance analysis of space complexity for a different number of student data collected from the dataset. The observed results confirmed that the CMN-GCCDSC technique considerably provides the better results when compared to the existing GA based decision tree classifier [1]. As the numbers of student data were increased, the space complexity gets increased.

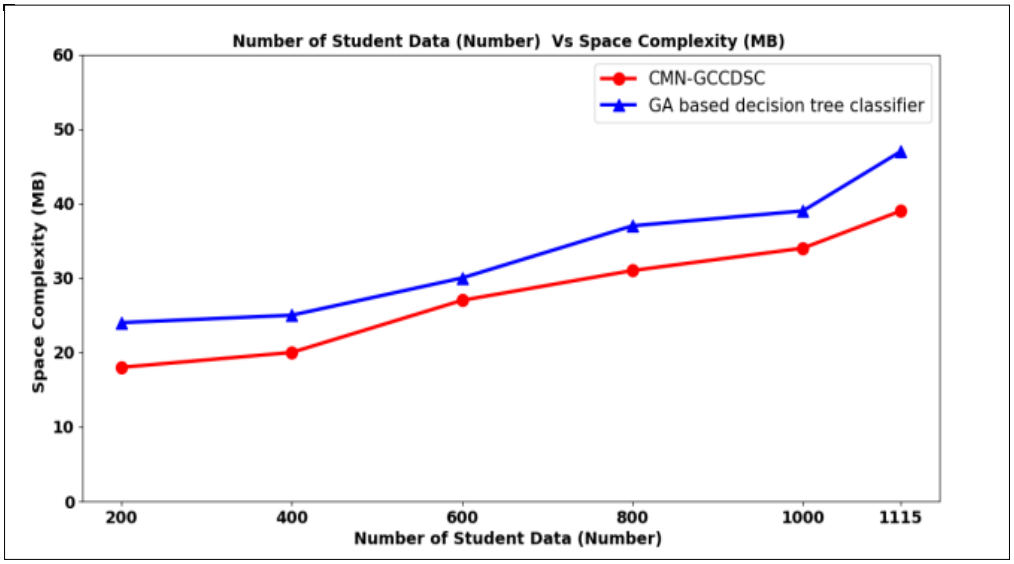

Fig. 3. Space Complexity

Figure 3provides the performance analysis of space complexity for a different number of student data collected from the dataset. The average of comparison results shows that the space complexity of CMN-GCCDSC technique is significantly minimized by 17% when compared to the existing GA [1]. Thus CMN-GCCDSC preprocessed the data by applying a Canberra Match Data Normalization technique to minimize the space complexity.

4.2 Time Complexity

Time complexity is measuredas the amount of time consumed for categorizing the student academic performance. The formula for calculating time complexity is given in eq (8).

$$TC = Number\ of\ student\ data * Time\ consumed\ by\ one\ data \tag{8}$$

Where, TC denotes a time complexity measured in milliseconds (ms).

Table 4. Performance results of time complexity

Number of Student Data (Number)	Time Complexity (ms)	
	GA based decision tree classifier	CMN-GCCDSC
200	15	12
400	16	14
600	21	18
800	26	22
1000	28	24
1115	36	31

Table 4 indicates the performance outcomes of time complexity versus the number of student data. The observed results confirmed that the CMN-GCCDSC technique considerably provides the better results when compared to the existing GA based decision tree classifier [1].

Figure 4 indicates the performance outcomes of time complexity versus the number of student data. The obtained performance results show thetime complexity is reducedusing CMN-GCCDSCby 15% when compared to existing GA [1]. This improvement is achieved due to application of Canberra Match Data Normalization based preprocessing forremoving the repeated or duplicate data from the dataset.

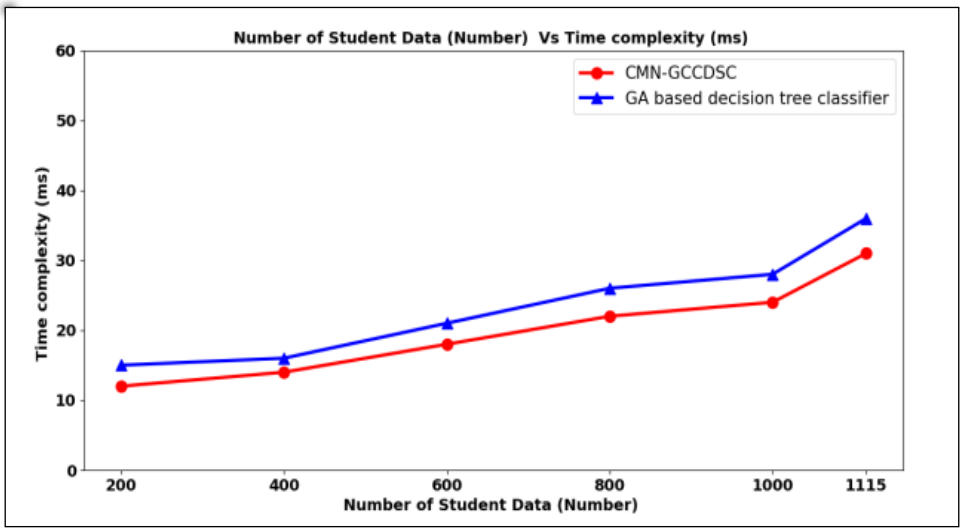

Fig. 4. Time Complexity

4.3 Accuracy

Accuracy refers to the ratio of the number of correctly classifiedstudent data to the total number of student data. The formula for calculating accuracy is given in eq (9).

$$Accuracy = \frac{Number\ of\ student\ data\ more\ accurately\ classified}{Numberof\ student\ data} * 100 \quad (9)$$

Table 5. Performance results of accuracy

Number of Student Data (Number)	Accuracy (%)	
	GA based decision tree classifier	CMN-GCCDSC
200	78	81
400	79	83
600	80	85
800	81	87
1000	82	88
1115	83	90

Table 5 indicates the performance outcomes of accuracy versus the number of student data. The observed results confirmed that the CMN-GCCDSC technique considerably provides the better results when compared to the existing GA based decision tree classifier [1].

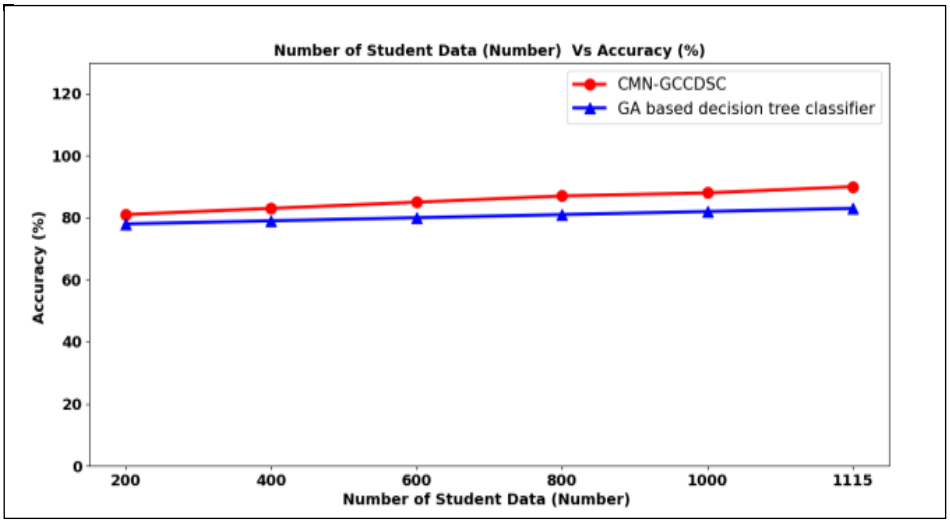

Fig. 5. Accuracy

Figure 5 illustrates the accuracy using two different techniques: CMN-GCCDSC, and GA-based decision tree classifier [1], based on the number of student data. The results indicate that the CMN-GCCDSC technique improves accuracy by 6% compared to GA [1].

4.4 Sensitivity

Sensitivity measures the proportion of correctly classified positive data out of the total number of data. It is expressed as a percentage (%). The formula for calculating sensitivity is given in eq (10).

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ Nagative} * 100 \tag{10}$$

Table 6. Performance results of sensitivity

Number of Student Data (Number)	Sensitivity (%)	
	GA based decision tree classifier	CMN-GCCDSC
200	77	80
400	78	81
600	79	82
800	80	84
1000	81	85
1115	82	87

Table 6 indicates the performance outcomes of sensitivity versus the number of student data. The observed results confirmed that the CMN-GCCDSC technique considerably provides the better results when compared to the existing GA based decision tree classifier [1].

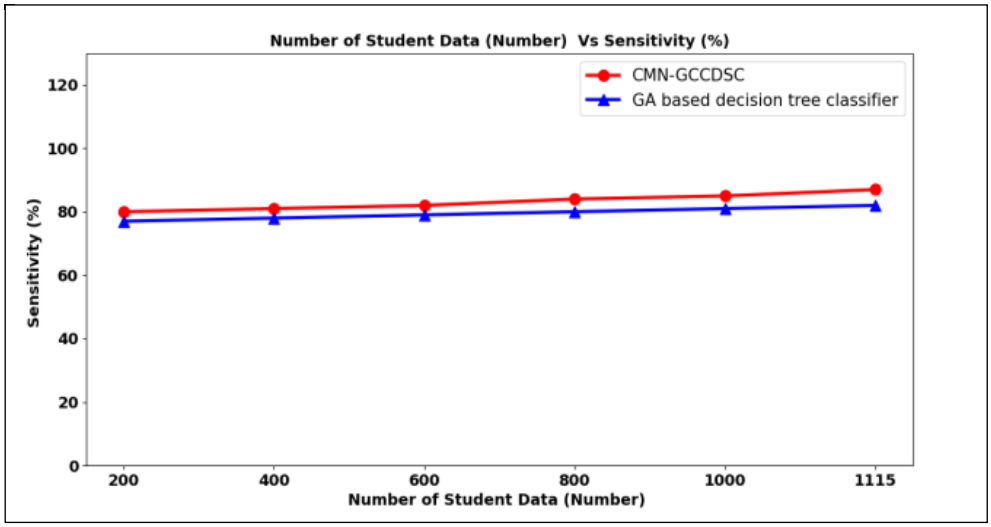

Fig. 6. Sensitivity

Figure 6, depicts the sensitivity analysis performance of two methods: CMN-GCCDSC, and GA-based decision tree classifier [1]. The canonical correlative decision stump classifier constructs decision rules based on correlation measures between the student academic performance and mobile addiction. These phases enhance the sensitivity of the CMN-GCCDSC. The overall results indicate that the CMN-GCCDSC method achieves a 5% improvement in overall sensitivity compared to GA [1].

4.5 Specificity

Specificity, also known as a classification metric, is defined as the ratio of the number of student data that are incorrectly classified to the total number of student data. It is measured as a percentage (%). The formula for calculating specificity is given in eq (11).

$$Specificity = \frac{Number\ of\ student\ data\ \ incorrectly\ classified}{Numberofstudent\ data} * 100 \tag{11}$$

Table 7. Performance results of specificity

Number of Student Data (Number)	Specificity (%)	
	GA based decision tree classifier	CMN-GCCDSC
200	15	10
400	14	9
600	13	8
800	12	7
1000	11	6
1115	9	5

Table 7 indicates the performance outcomes of specificity versus the number of student data. The observed results confirmed that the CMN-GCCDSC technique considerably provides the better results when compared to the existing GA based decision tree classifier [1].

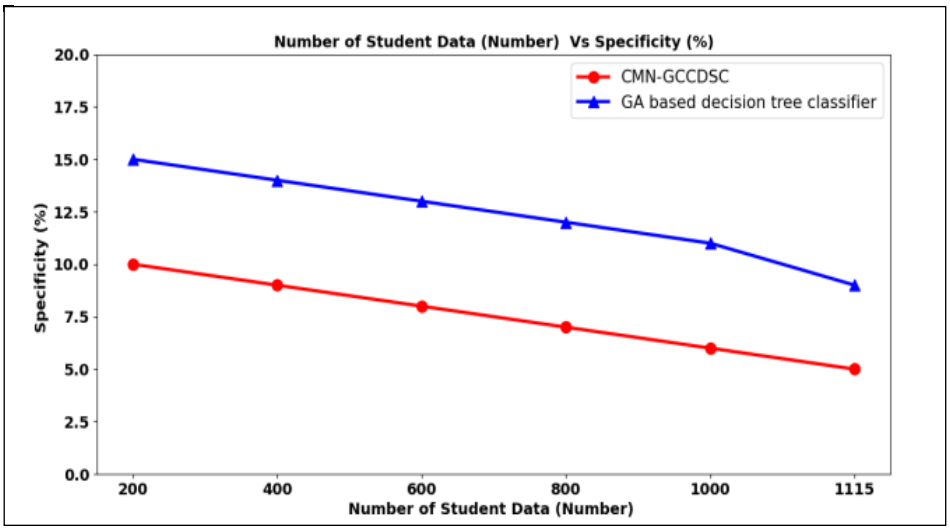

Fig. 7. Specificity

Figure 7 illustrates the graphical representation of specificity versus the number of student data points. The overall comparison results demonstrate that the CMN-GCCDSC technique minimizes the specificity by 40% compared to GA [1]. This improvement is achieved by the effective categorization of student data by setting a threshold value and minimizing incorrect data classification.

5. Conclusion

Smartphone addiction and its harmful effects on academic performance have received attention from both students and practitioners. A CMN-GCCDSC is developed to classify academic performance of students. In CNERBC Technique, Canberra Match Data Normalization Process is performed to pre-process the data for removing the repeated data from database. Every data is checked with another data to find repeated one for minimizing the dimensionality. By this way, student academic performance prediction is carried out in efficient manner. A comprehensive experimental assessment is conducted with metrics, including space complexity, time complexity, accuracy, sensitivity, and specificity. It helps to enhance the accuracy in predicting student academic performance. The time complexity is reduced and thereby improves the run time efficiency of the algorithm. As a result, students' academic performances were predicted using different attributes, algorithms, and techniques. The results confirm that machine learning algorithms can be used to predict students' academic performance.

By using mobile addiction attributes, student academic performance is classified such as Marginal, Below Average, Average, Above Average, Good and Excellent. It also provides higher accuracy by 6%, sensitivity by5% and minimized the specificity by 40%, space complexity by 17% as well as time complexity by 15% when compared to the existing GA based decision tree classifier method. The CMN-GCCDSC technique is an effective solution that addresses the limitations of Genetic Algorithm (GA)-based decision tree classifiers. By combining the Decision Stump Classifier (DSC) approach with Generalized Canonical Correlation (GCC), the most important feature to consider for academic prediction among students were selected, ultimately reducing the dimensionality of the dataset, and improving classifier performance. With higher accuracy rates achieved, this technique can help identify at-risk students early and discover hidden trends and patterns in student performance, leading to improved academic outcomes with additional support from institutions and faculties.

6. Future Work

The proposed method, which focuses on preprocessing and classification, has shown to be highly effective in predicting student academic performance, with increased accuracy, reduced time and space complexity, and improved sensitivity and specificity. To further enhance this approach, future research can explore more refined attribute selection and classification. Future work should include additional features and attributes that were not covered in this study. Furthermore, more algorithms can be utilized to optimize predictive results. By analyzing a larger number of Machine Learning algorithms and incorporating additional features, the prediction of student academic performance will be more accurate.

References

[1] S. Hussain and M. Q. Khan, "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of data science*, vol. 10, pp. 637-655, 2023.

[2] S. Rajendran, S. Chamundeswari, and A. A. Sinha, "Predicting the academic performance of middle-and high-school students using machine learning algorithms," *Social Sciences & Humanities Open*, vol. 6, p. 100357, 2022.

[3] R. R. Ahmed, F. Salman, S. A. Malik, D. Streimikiene, R. H. Soomro, and M. H. Pahi, "Smartphone use and academic performance of university students: a mediation and moderation analysis," *Sustainability*, vol. 12, p. 439, 2020.

[4] J. C. Wang, C.-Y. Hsieh, and S.-H. Kung, "The impact of smartphone use on learning effectiveness: A case study of primary school students," *Education and Information Technologies*, vol. 28, pp. 6287-6320, 2023.

[5] B. Rathakrishnan, S. S. Bikar Singh, M. R. Kamaluddin, A. Yahaya, M. A. Mohd Nasir, F. Ibrahim, *et al.*, "Smartphone addiction and sleep quality on academic performance of university students: An exploratory research," *International journal of environmental research and public health*, vol. 18, p. 8291, 2021.

[6] S. Li and T. Liu, "Performance prediction for higher education students using deep learning," *Complexity*, vol. 2021, pp. 1-10, 2021.

[7] C. F. Rodr ́guez-Hern ́andez, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100018, 2021.

[8] B. K. Yousafzai, S. A. Khan, T. Rahman, I. Khan, I. Ullah, A. Ur Rehman, *et al.*, "Student-performulator: Student academic performance using hybrid deep neural network," *Sustainability*, vol. 13, p. 9775, 2021.

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

[10] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827-203844, 2020.

[11] L. Zhao, K. Chen, J. Song, X. Zhu, J. Sun, B. Caulfield, *et al.*, "Academic performance prediction based on multisource, multifeature behavioral data," *IEEE Access*, vol. 9, pp. 5453-5465, 2020.

[12] E. Alhazmi and A. Sheneamer, "Early Predicting of Students Performance in Higher Education," *IEEE Access*, vol. 11, pp. 27579-27589, 2023.

[13] A. Rafique, M. S. Khan, M. H. Jamal, M. Tasadduq, F. Rustam, E. Lee, *et al.*, "Integrating learning analytics and collaborative learning for improving student's academic performance," *IEEE Access*, vol. 9, pp. 167812-167826, 2021.

[14] C. Cui, J. Zong, Y. Ma, X. Wang, L. Guo, M. Chen, *et al.*, "Tri-Branch Convolutional Neural Networks for Top-k Focused Academic Performance Prediction," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[15] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558-19571, 2022.

[16] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human behavior*, vol. 104, p. 106189, 2020.

[17] F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, "A deep learning classification framework for early prediction of team-based academic performance," *Applied Soft Computing*, vol. 106, p. 107355, 2021.

[18] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Computers in Human Behavior*, vol. 98, pp. 166-173, 2019.

[19] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of medical systems*, vol. 43, pp. 1-15, 2019.

[20] A. Akram, C. Fu, Y. Li, M. Y. Javed, R. Lin, Y. Jiang, *et al.*, "Predicting students' academic procrastination in blended learning course using homework submission data," *Ieee Access*, vol. 7, pp. 102487-102498, 2019.

[21] D. Sun, R. Luo, Q. Guo, J. Xie, H. Liu, S. Lyu, *et al.*, "A University Student Performance Prediction Model and Experiment Based on Multi-Feature Fusion and Attention Mechanism," *IEEE Access*, 2023.

[22] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a web-based prediction system for students' academic performance," *Data*, vol. 7, p. 21, 2022.

[23] M. Kumar, G. Mehta, N. Nayar, and A. Sharma, "EMT: Ensemble meta-based tree model for predicting student performance in academics," in *IOP Conference Series: Materials Science and Engineering*, 2021, p. 012062.

[24] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, *et al.*, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608-95621, 2021.

[25] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach," *IEEE Access*, vol. 11, pp. 136091-136108, 2023.

Authors' Profiles

R. Ruth Belina is a Research Scholar in the Department of Computer Science, Holy Cross College, Tiruchirappalli, Tamil Nadu, India. She has 16 years of teaching experience. She has presented research articles at various International Conferences. Her areas of interest are Computer Networks, Software Engineering and Machine Learning.

T. Lucia Agnes Beena is working as an Assistant Professor in the Department of Computer Science, Holy Cross College, Tiruchirappalli, Tamil Nadu, India. She has 21 years of teaching experience and 8 years of research experience. She has published several research articles in Scopus indexed Journals. She authored one book and edited one book. She has published more than 12 chapters with CRC, Wiley and IET publishers. Her areas of interest are Cloud Computing, Psychology of Computer Programming and Machine Learning.

Charles Savarimuthu, who specializes in data mining and big data, has 24 years of teaching and 12 years of research experience. He has a strong record of scholarly publications, with a Scopus/WoS score of 7/3 and h-indexes of 4 and 2 on Google Scholar and Web of Science, respectively. He is working as a senior lecturer in the Department of Information Technology at Mussanah University of Technology and Applied Sciences, Oman.

How to cite this paper: R. Ruth Belina, T. Lucia Agnes Beena, Charles Savarimuthu, "Canberra Match Normalization-Enhanced Decision Stump Classifier for Predicting Academic Performance in the Context of Smartphone Addiction", International Journal of Modern Education and Computer Science(IJMECS), Vol.17, No.1, pp. 59-71, 2025. DOI:10.5815/ijmecs.2025.01.05