Optimizing the k value in the k Nearest Neighbor Algorithm for Academic Prediction of Working Students

Rakhi Paul
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh

Mithun Das
Department of Computer Science and Engineering
BGC Trust University Bangladesh
Chattogram, Bangladesh

ABSTRACT
This study highlights the dedication of the institution to cultivating scholastic distinction, specifically among pupils who are managing the concurrent obligations of employment and education. This study endeavors to establish a resilient framework for forecasting and categorizing scholastic achievement by utilizing the K nearest neighbor algorithm and optimizing the value of k via 5-fold cross-validation. By including three distinct label classes—satisfactory, very satisfactory, and with honors—the model's predictive capability is enhanced, providing more nuanced insights into the academic accomplishments of students. Significantly, the accuracy rate of 85.71% achieved in determining k=3 highlights the effectiveness of the proposed methodology in accurately predicting academic outcomes. The results of this study have substantial ramifications for policymakers in academia, providing them with empirical data that can guide the creation of individualized interventions and policies that promote the comprehensive growth of working students. Furthermore, this study makes a valuable contribution to the wider academic conversation regarding predictive analytics in the field of education by providing a balanced analysis of the complex elements that impact student achievement. This study sheds light on potential areas for targeted support and intervention by examining the relationship between employment obligations and academic achievement. As a result, it promotes a more favorable learning environment for every student. Moreover, the incorporation of sophisticated machine learning methods highlights the establishment's dedication to harnessing state-of-the-art approaches to improve academic achievements. The high accuracy rate of the predictive model serves as evidence of the K nearest neighbor algorithm's effectiveness in capturing the intricate dynamics that are intrinsic to the academic trajectories of students. In general, this study signifies a substantial progression in the direction of enhancing educational policies and practices to more effectively address the requirements of a heterogeneous student population, thereby furthering the overarching objective of scholastic distinction and student achievement.

General Terms
K Nearest Neighbor Algorithm, Resilient Framework, 5-fold Cross-validation, Machine Learning Methods, Concurrent Obligations.

Keywords
Machine Learning, k Nearest Neighbor Algorithm, Predictive Analytics, Academic Performance, Predictive Model.

1. INTRODUCTION
An educational institution's main goal is to improve students' academic performance by offering them a high-quality education [1]. This study attempts to categorize and forecast students' academic performance, with a focus on working students since they have a greater workload than non-working students. Academic officials can take the outcomes of this classification and prediction into account when making choices about working students. One technique for obtaining previously undiscovered information from big databases is data mining. It aids in the analysis of potential patterns and trends, empowering institutions to make wise choices. To provide insightful knowledge and conclusions, the data analysis process entails evaluating, cleansing, and modeling data [2]. In data mining, one popular classification method is the K-nearest neighbor algorithm. It is a nonparametric pattern recognition algorithm that has been widely used in various fields due to its simplicity, effectiveness, and intuitiveness [3]. This algorithm is simple yet effective for predicting class labels of a query based on surrounding information [3]. However, the algorithm has a weakness in selecting the value of k. Selecting the value of k in the kNN algorithm is crucial. The value of "k," which represents the number of neighbors based on distance metrics, greatly influences classification [4]. Various techniques have been proposed for selecting the value of k, such as cross-validation and heuristics. The value of k should not be a multiple of the number of classes to avoid ties. Additionally, a large k value reduces the effect of noise on classification and makes class boundaries less clear, while a small k value makes classification results more susceptible to noise [5]. K-fold cross-validation is a well-liked approach for assessing a classification dataset's performance [6]. The process of K-fold cross-validation entails splitting the input data set into K folds, each of which is utilized as a testing set once. Juggling work obligations and academic obligations presents special difficulties for students who are employed. For these pupils, increasing the k value in the k Nearest Neighbor Algorithm offers a chance to improve the precision of academic forecasts [13].

However, determining the optimal k value requires a systematic investigation given the complexity of academic and work-related factors. What is the optimal k value in the k Nearest Neighbor Algorithm for accurately predicting academic performance among working students?

In the pursuit of enhancing academic forecasting methodologies amidst the dynamic landscape of higher education, this research endeavors to delve into the intricate interplay between academic performance and employment engagements among students [14]. By examining the application of the k Nearest Neighbor Algorithm, this study aims to not only review existing literature but also address pertinent challenges faced by working students. Through

meticulous analysis and methodological exploration, the objectives of this research are delineated as follows:

- To review previous studies on the application of the k Nearest Neighbor Algorithm in academic prediction.
- To identify and analyze the specific challenges faced by working students in managing their academic and work commitments.
- To explore existing methods and techniques for optimizing the k value in the k Nearest Neighbor Algorithm.
- To describe the dataset comprising academic performance indicators, working hours, and other relevant attributes.
- To perform preprocessing steps including data cleaning, normalization, and feature selection to prepare the dataset for analysis.
- To implement the k Nearest Neighbor Algorithm and develop a framework for optimizing the k value.

2. RESEARCH METHODOLOGY

The academic performance data of students from the Department of Computer Science who studied while employed from the 2018 cohort to the 2020 cohort is the data source utilized in this study. Training and testing data will be the two categories into which the data will be split.

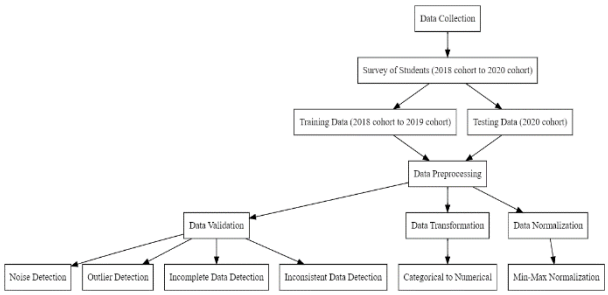

Fig. 1. Research Stages and Data Processing Flow

2.1 Data Collection

The survey, which was conducted among students from the 2018 cohort to the 2020 cohort at X University Bangladesh's Department of Computer Science and Engineering, is the source of the data. Training and testing data will be the two categories into which the data will be split [11]. Thirty students from the 2020 cohort will make up the testing data, while students from the 2018 cohort to the 2020 cohort will be used for training. Student ID, name, GPA for each semester (up to semester 5), weekly working hours, and monthly income are among the attributes that are gathered [12].

2.2 Initial Data Processing

Data preprocessing is a crucial stage in designing a classification model [7]. Initial data processing is performed to obtain clean data that is free from noise or outliers. Several steps are involved:

- **Data validation:** Data validation is performed to identify any noise, outliers, incomplete, or inconsistent data.
- **Data transformation:** In this research, there are several categorical data. These data are transformed into numerical data.
- **Data normalization:** Min-max normalization is used in this research. Min-max normalization performs a linear transformation on the original data [8], aiming to obtain balanced attribute values or to produce data within a specific range. Unbalanced value ranges for each attribute can affect the quality of data mining results.

2.3 Proposed Methodology

This research will be conducted using the K-Nearest Neighbor method to classify and predict data from students at the Department of Computer Science and Engineering, X University Bangladesh, who study while working. For optimizing the value of k, k-fold cross-validation will be utilized. Cross-validation is one of the most commonly used methods for estimating the complexity of a model [9].

2.4 Model Testing

In model testing, optimization of the k value will be performed first using 5-fold cross-validation. Subsequently, this value will be used as the k value in the kNN process. The next step involves performance testing to assess the accuracy of the classification. The final step is to conduct prediction testing on the testing data [15].

2.5 Evaluation and Validation

Validation is a crucial stage in modelling to assess the reliability of the model for decision-making [10]. For the evaluation and validation of classification results, the confusion matrix will be examined. Meanwhile, prediction results will be presented in the form of a prediction result table.

3. RESULTS AND DISCUSSION

3.1 Training Data

The training data consists of students from the 2018 cohort to the 2020 cohort. The training data used can be seen in Table 1.

No.	ID	Name	GPA Semester 1	GPA Semester 2	GPA Semester 3	GPA Semester 4	GPA Semester 5	Working Hours	Attendance	GPA	Academic Performance
1	1857201049	Aisha Ahmed	2.85	2.87	2.92	2.68	2.44	2	1	2.75	Satisfactory
2	1955201005	Yusuf Ali	3.60	3.86	3.75	3.63	3.50	5	4	3.67	With Praise
3	1955201101	Fatima Hassan	2.40	2.40	1.50	2.27	2.07	5	2	2.13	Satisfactory
4	2055201067	Muhammad Khan	3.05	3.04	3.00	2.50	3.60	5	4	3.04	Very Satisfactory

Below are the transformations for the weekly working hours and monthly income data as shown in Tables 2 and 3 respectively:

Table 2: Transformation for Weekly Working Hours

Working Hours Range	Transformation
1 hour to 10 hours per week	1
11 hours to 20 hours per week	2
21 hours to 30 hours per week	3
31 hours to 40 hours per week	4
More than 40 hours per week	5

Table 3. Transformation of Income Data

Income Range	Transformation
BDT 10,000 to BDT 15,000	1
BDT 15,000 to BDT 25,000	2
BDT 25,000 to BDT 35,000	3
More than BDT 35,000	4

Tables 2 and 3 demonstrate the transformed data for weekly working hours and monthly income, which will be used in the classification and prediction process.

3.2 Testing Data

The testing data consists of students from the 2020 cohort, totaling 30 data points. The testing data can be seen in the following table:

Table 4: Testing Data

No	ID	Name	Semester Grade 1	Semester Grade 2	Semester Grade 3	Weekly Working Hours	Income
1	2055201002	Nazma Akter	3.50	3.65	3.88	1	1
2	2055201003	Rezaul Karim	3.37	3.26	3.63	2	1
3	2055201008	Kamal Hossain	2.74	3.6	3.5	5	1
4	2055201009	Abdullah Hasan	3.58	3.26	3.38	1	1
...	.....	.....	.....	.....	.....	.....	.....
30	2057201069	Yasmin Chowdhury	3.58	3.26	3.38	1	1

Table 4 presents the testing data for students from the 2019 cohort, including their student ID, name, GPA for each semester (Semesters 1 to 3), weekly working hours, and monthly income.

3.3 Data Normalization

Normalization is one of the preprocessing techniques used to handle a range of attribute values so that data is distributed on the scale same [7]. In this research, the normalization process was carried out using the Min method Max with a value range between 0 and 1. Results of the normalization of data training and testing data can be seen in Table 5 and Table 6 below.

Table 5. Training Data Normalization

No	ID	GPA Semester 1	GPA Semester 2	GPA Semester 3	GPA Semester 4	GPA Semester 5	Working Hours	Monthly Income	Academic Performance GPA
1	1557201049	0.28	0.34	0.63	0.23	0.61	0	0	Satisfactory
2	1655201005	0.75	0.99	1	0.78	0.87	1	1	With Honors
3	1655201101	0	0.03	0	0	0.51	1	0.33	Satisfactory
4	1755201067	0.40	0.45	0.66	0.13	0.9	1	1	Very Satisfactory
5	1755201068	0.71	0.34	0.72	0.34	0.78	1	0.66	Very Satisfactory
...	...	...	...	...	...	...	...	...	...
35	1857201080	0.37	0.41	0.74	0.61	0.85	0	0	Very Satisfactory

Table 5 represents the normalized training data. The data has been scaled between 0 and 1 for each attribute (GPA, Semester 1, GPA Semester 2, GPA Semester 3, GPA Semester 4, GPA Semester 5, Working Hours, Monthly Income) to ensure uniformity and aid in the classification and prediction process. Additionally, the academic performance GPA column reflects the corresponding performance level.

Table 6. Normalized Testing Data

No	ID	Academic Performance Index Semester 1	Academic Performance Index Semester 2	Academic Performance Index Semester 3	Weekly Working Hours	Monthly Income
1	1955201002	0.6	0.79	0.86	0	0
2	1955201003	0.5	0.18	0.56	0.25	0
3	1955201008	0	0.71	0.41	1	0
...	...	...	...	...	...	...
30	1957201061	1	0.73	0.56	0.5	0

This table represents the normalized testing data. The data has been scaled between 0 and 1 for each attribute (Academic Performance Index for Semester 1 to 3, Weekly Working Hours, and Monthly Income) to ensure uniformity and aid in the classification and prediction process.

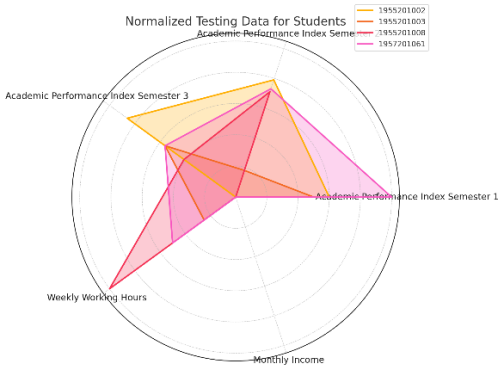

Fig. 2. Normalized Testing Data for Students

Figure 2 is used to illustrate a radar chart that shows the results of the normalized testing data of the chosen students. A radar chart has been selected as it makes it easier to understand how the academic performance indices for three semesters, as well as weekly working hours and monthly income, vary. The radar chart has a line for each student, with the normalized scores of each student demonstrated across the five dimensions.

3.4 Optimization of k Value

The optimization of the k value is performed using the k-fold validation algorithm. The number of folds used is 5. The process of the k-fold validation algorithm can be seen in the following flowchart (Figure 3).

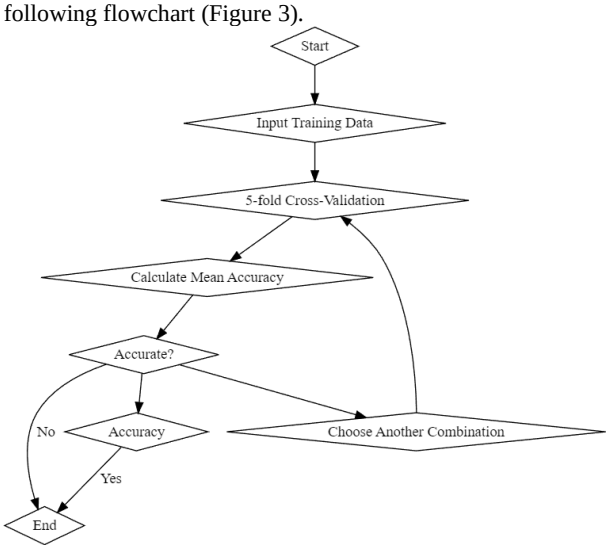

Fig. 3. Training and Validation Flowchart with 5-fold Cross-Validation

From the results of 5-fold cross-validation, an accuracy value of 0.55 is obtained with a k value of 3 (k=3). This value will be subsequently used as the k value for the kNN process. Below is the result of 5-fold cross-validation (Figure 4).

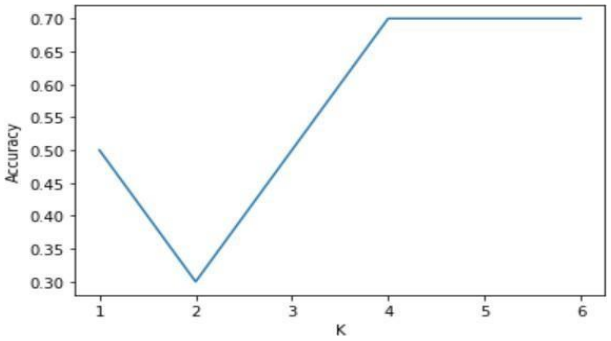

Fig. 4. Graph of Accuracy Level for k Values

The precision of the k-nearest Neighbors (kNN) algorithm is plotted as a function of the hyperparameter 'k' in Figure 3, which indicates the number of neighbors to take into account for forecasting purposes. The graph's y-axis displays the associated accuracy values, while the graph's x-axis displays the values of 'k', which range from 1 to 6.

The graph shows us the following observations:

- When k=1, the accuracy is just above 0.35, which is relatively low.
- As k increases to 2, the accuracy drops to its lowest point on the graph, which is around 0.33.
- At k=3, the accuracy increases significantly to approximately 0.55, which is the value you've mentioned.
- The accuracy continues to rise as k increases to 4, reaching around 0.65.
- For k=5 and k=6, the accuracy remains relatively stable at the peak value of approximately 0.65.

Based on this graph, it appears that k=3 provides a substantial improvement over k=1 and k=2, but it is not the optimal value for accuracy. The graph suggests that k=4, k=5, and k=6 yield even higher accuracy. However, if for some reason k=3 has been chosen as the value for subsequent kNN processes, it could be due to considerations not visible in the graph, such as a trade-off between accuracy and model complexity, computational efficiency, or variance in the results.

3.5 kNN Process

The kNN process is performed using RapidMiner. The first step is to assess the accuracy level generated from 5- fold cross-validation with a k value of 3. The stages of cross-validation with RapidMiner can be seen in the following figure, Figure 5.

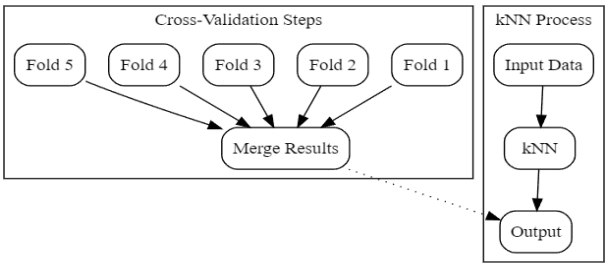

Fig. 5. RapidMiner Steps to Assess Accuracy Level

From the above 5-fold cross-validation results with k=3, the accuracy level obtained is 85.71% with the following Confusion Matrix (Table 7).

Table 7. Confusion Matrix

	True Satisfactory	**True with Praise**	**True Very Satisfactory**	**Class Precision**
Predicted Satisfactory	1	0	0	100%
Predicted with Praise	0	10	1	90.91%
Predicted Very Satisfactory	3	1	19	82.61%
Class Recall	25%	90.91%	95%	

The confusion matrix, which is a vital tool for assessing how well classification models work, is shown in the table. It is set up so that rows stand for the anticipated classes and columns for the actual classes.

Here's a breakdown:

- The first row indicates that one instance predicted as "Satisfactory" was correct.
- In the second row, out of 11 instances predicted as "With Praise," 10 were accurate.
- The third row shows that out of 23 instances predicted as "Very Satisfactory," 19 were accurate.

For each class, the precision is shown as a percentage: "Very Satisfactory" is 82.61% accurate, "With Praise" is 90.91% accurate, and "Satisfactory" is 100% accurate. The "Class Recall" row also displays each class's recall percentage; "Very Satisfactory" is at 95%, "With Praise" is at 90.91%, and "Satisfactory" is at 25%.

3.6 Prediction

Prediction testing is conducted using RapidMiner. In this prediction stage, the testing data will be evaluated with the training data to observe the prediction results regarding academic performance with a k value of 3. The flow diagram of this process can be seen in Figure 6.

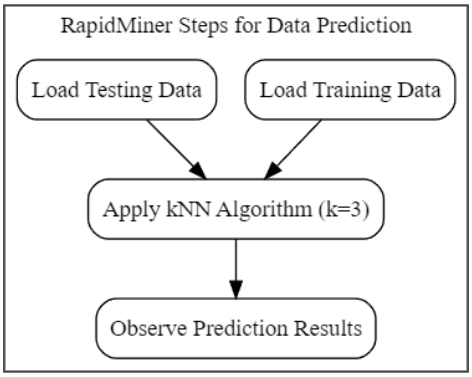

Fig 6. RapidMiner Steps for Data Prediction

The results of the prediction test can be seen in Table 8.

Table 8. Prediction Test Results for Testing Data

No	Student ID	Confidence (Satisfactory)	Confidence (with Praise)	Confidence (Very Satisfactory)	Prediction (Academic Performance)
1	1955201002	0	0.602	0.397	With Praise
2	1955201003	0.395	0	0.604	Very Satisfactory
3	1955201008	0.607	0	0.392	Satisfactory
4	1955201009	0.396	0	0.603	Very Satisfactory
...	...	...	...	...	...
30	1957201069	0.396	0	0.603	Very Satisfactory

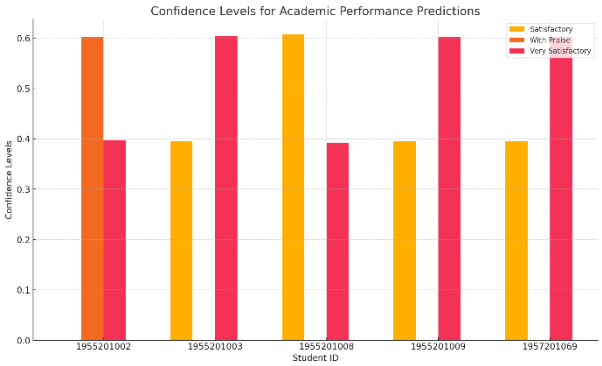

Fig 7. Confidence Levels for Academic Performance Predictions

Figure 7 reveals the confidence levels of the student's predicted achievement, in Satisfactory, With Praise, and Very Satisfactory. The figure also represents the distribution of students' predicted scores that support a particular level of confidence in the prediction, as well as the distribution of students' predicted scores that support a different level of confidence. The purpose of this visualization is to show that the distribution of predicted achievement scores generally aligns with confidence levels.

4. CONCLUSION

The study, which focused on students who hold parallel jobs, used data from the Department of Computer Science and Engineering at X University, Bangladesh. The testing dataset only contained students from the 2020 cohort, whereas the training dataset included students from the academic cohorts of 2018 to 2020. Three performance labels were applied to these students: extremely satisfactory, with acclaim, and satisfactory. Even though optimization using 5-fold cross-validation and a k value of 3 resulted in an accuracy rate of 0.55, the study had certain drawbacks. The size and representativeness of the dataset may be one such constraint. The statistics might not accurately reflect the variety of academic achievements among students or the specifics of their work-study schedules. Furthermore, the study might have missed some outside variables that affect academic achievement, such as socioeconomic backgrounds or pressures from the workplace. Future research could use several approaches to improve the study's applicability and robustness. First off, broadening the dataset to include a wider variety of students from various departments or universities may offer a more comprehensive understanding of the connection between academic achievement and work-study dynamics. Furthermore, adding more characteristics or variables—like the nature of the work, the duties performed, or the number of hours worked—might provide a more detailed understanding of the precise elements influencing student success. Moreover, utilizing cutting-edge machine learning methods or investigating different algorithms outside of kNN may enhance prediction precision and model generalization. Methods like deep learning or ensemble learning could be investigated to find intricate patterns in the data and improve prediction abilities. Finally, combining quantitative analysis with qualitative research—such as surveys or interviews—could lead to a deeper comprehension of students' perspectives on the relationship between work-study balance and academic achievement. Combining qualitative and quantitative research findings could result in more thorough and practical recommendations for educational

institutions and legislators who want to successfully assist working students.

5. REFERENCES

[1] F. Okubo, "A Neural Network Approach for Students ' Performance Prediction," no. March, pp. 5–7, 2017, doi: 10.1145/3027385.3029479.

[2] L. Mahmoud and A. Zohair, "Prediction of Student ' s performance by modelling small dataset size," 2019.

[3] A. Olawoyin, Y. Chen, A. Olawoyin, and Y. Chen, "ScienceDirect ScienceDirect ScienceDirect Predicting the Future with Artificial Neural Network Predicting the Future with Artificial Neural Network," Procedia Comput. Sci., vol. 140, pp. 383–392, 2018, doi: 10.1016/j.procs.2018.10.300.

[4] A. A. Aryaguna and D. O. Anggriawan, "Identifikasi Jenis Gangguan Pada Jaringan Distribusi Menggunakan Metode Artificial Neural Network," no. April, pp. 27–35, 2021.

[5] Abhijit Pathak, Arnab Chakraborty, Minhajur Rahaman, Taiyaba Shadaka Rafa, and Ummay Nayema, "Enhanced Counterfeit Detection of Bangladesh Currency through Convolutional Neural Networks: A Deep Learning Approach", IJIRCST, vol. 12, no. 2, pp. 10–20, Mar. 2024.

[6] A. Çetinkaya and Ö. K. Baykan, "Prediction of middle school students' programming talent using artificial neural networks," Eng. Sci. Technol. an Int. J., no. xxxx, 2020, doi: 10.1016/j.jestch.2020.07.005.

[7] E. Bahadır, "Prediction of Prospective Mathematics Teachers' Academic Success in Entering Graduate Education by Using Back- propagation Neural Network," J. Educ. Train. Stud., vol. 4, no. 5, pp. 113–122, 2016, doi: 10.11114/jets.v4i5.1321.

[8] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," 2019, doi: doi.org/10.1007/s10639-019-10053.

[9] A. Ali and N. Senan, "The Effect of Normalization in Violence Video Classification Performance," IOP Conf. Ser. Mater. Sci. Eng., vol. 226, no. 1, 2017, doi: 10.1088/1757-899X/226/1/012082.

[10] Streamlining Visitor Appointments: Automated Scheduling System for BGC Trust University Bangladesh. (2023). International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets44679.

[11] G. Jiang and W. Wang, "Error estimation based on variance analysis of k-fold cross-validation," Pattern Recognit., vol. 69, pp. 94– 106, 2017, doi: 10.1016/j.patcog.2017.03.025.

[12] Ali and N. Senan, "The Effect of Normalization in Violence Video Classification Performance," IOP Conf. Ser. Mater. Sci. Eng., vol. 226, no. 1, 2017, doi: 10.1088/1757-899X/226/1/012082.

[13] Hossen, H., Shuvon, M. S. S., Barsha, J. B., Chy, A. A., & Pathak, A. (2023). Ultimate cricket experience: Dynamic web app for a real-time scoring system in university cricket. World Journal of Advanced Research and Reviews, 19(02), 1269–1280. DOI: 10.30574/wjarr.2023.19.2.1721.

[14] S. Eker, E. Rovenskaya, S. Langan, and M. Obersteiner, "Model validation: A bibliometric analysis of the literature," Environ. Model. Softw., vol. 117, no. December 2018, pp. 43–54, 2019, doi: 10.1016/j.envsoft.2019.03.009.

[15] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," Expert Syst. Appl., vol. 106, pp. 252–262, 2018, doi: 10.1016/j.eswa.2018.04.008.