Analyzing the Correlation Between Student Club Participation, Co-Curricular Activities, Extra-curricular Activities and Their Influence on Academic Success and Employment Outcomes Among University Students

Md. Maruf Hossen
Student ID: 0242220005101373
Computer Science & Engineering
Daffodil International University
Birulia, Savar, Dhaka
maruf1015hossain@gmail.com

Md. Jubair Hossain
Student ID: 0242220005101395
Computer Science & Engineering
Daffodil International University
Birulia, Savar, Dhaka
hossain22205101395@diu.edu.bd

Md. Ismail Hossain
Student ID: 0242220005101001
Computer Science & Engineering
Daffodil International University
Birulia, Savar, Dhaka
ismilehossain211@gmail.com

***Abstract*—This study looks at how joining student clubs and participating in activities outside of class affects university students' grades and job opportunities. Most previous studies rely primarily on literature review, with only a limited number of papers conducting data-driven analysis. We collected data from 162 students at 32 universities and used a method called SMOTE to balance the data. Then, tested eight different computer models to analyze the information: Decision Tree, Naive Bayes, Random Forest, CatBoost, SVM, XGBoost, Logistic Regression, and k-Nearest Neighbors. The Decision Tree model worked best, with 71.1% accuracy and a matching F1 score, while Naive Bayes was great at spotting patterns, with an 80.95% recall rate. The findings show that students who take part in these activities tend to have better grades, develop more skills, build stronger connections with industries, and have better chances of landing jobs. However, the study has some limits: it relies on students' self-reported answers, uses artificial data to fill gaps, tests the data in only one way, and focuses on just one university. Future research should explore more places and try different ways of collecting data.**

***Keywords*—Student club involvement, co-curricular activities, extra-curricular participation, academic performance, employment outcomes, machine learning.**

I. Introduction

Involvement in student clubs, co-curricular, and extracurricular activities is widely recognized as a valuable aspect of university life. These activities, which extend beyond regular class studies, help students develop general skills like leadership, teamwork, networking and communication. Universities are increasingly concerned with developing students not only for academic success but also for their future careers, so it is relevant to understand how participation in these activities influences both academic success and employability. The primary aim of this study is to analyze whether involvement in student clubs, co-curricular, and extracurricular activities impacts the academic performance of students and equips them with skills to enhance their employability after graduation.

Previous research demonstrates these activities result in beneficial outcomes. Research by computer model, logistic regression, demonstrated 69% correlation between out-of-class activities and improved grades, and 99% success in classifying academic success [1]. Another research by Decision Tree model had an accuracy of 85% and F1 value of 84% in analyzing time spent on activities versus school performance [2]. A research on 148 high school students established that students engaged in at least a single activity had superior grades (3.456, on average) compared to students who were not (2.578), and the finding was supported by a confidence level of 99.995% [3]. For employment, it has been studied that employers are keen to find problem-solving, teamwork, and leadership skills, which are developed by students in these kinds of activities. Studies demonstrated students who are involved in club activities are offered employment after interview, and students who have leadership positions are usually offered jobs in bigger organizations, which are also of greater stability [4]. Internships, volunteering, or work-study programs teach employable skills [5]. Yet, students occasionally overestimate their skill set relative to what employers require, which requires further study. Numerous studies do either grades or employment, not both, and frequently use aged data or summaries rather than new student data. Moreover, studies assess participation in different ways, so it is difficult to compare them. This study bridges these gaps by using new data from university students and computer models to examine how clubs and activities influence both grades and employability.

With new evidence and sophisticated instrumentation, this study showcases how participation assists students in succeeding academically and in their professional lives. The results seek to offer valuable information for universities, students, and employers regarding the value of participating in extracurricular activities.

By leveraging machine learning and primary data, this study provides a comprehensive analysis of how student engagement in these activities fosters both academic success and career readiness. The findings aim to offer practical insights for universities, students, and employers, highlighting the tangible benefits of involvement in student clubs and activities beyond the classroom.

II. Literature Review

This review synthesizes sixteen studies on the impact of extracurricular and co-curricular activities on students' academic performance and employability. The evidence is discussed across four themes: academic outcomes, employability, Bangladeshi research, and study limitations. Methods vary, but few studies use real student datasets; most are surveys or literature reviews, limiting causal conclusions.

A. Impact on Academic Performance

Findings are mixed. A machine learning study of 850 students found a 69% correlation between activity participation and GPA [1], while a meta-analysis of 29 U.S. studies reported only modest achievement gains and no clear causal link [2]. Educator surveys note benefits for growth and grades [9], though other evidence shows habits and prior performance are stronger predictors [6]. Overall, activities may help but effects depend on context.

B. Effect on Employability

Activities develop soft skills valued in the job market. UK students recognized networking benefits, though their priorities differed from employers [7]. In Bangladesh, teamwork and communication explained 37.2% of employability variance [8]. Australian graduates valued extracurricular internships more than curricular ones [10]. Reviews also link activities to reduced risky behavior and better long-term career prospects [5]. Alignment with workplace needs determines impact.

C. Research in Bangladesh

Bangladeshi findings are inconsistent. Social or political activities often harm grades [11], [12], [16], while sports and cultural activities show positive effects [13], [16]. Other studies highlight stress reduction and skill acquisition but little academic improvement [14]. Most Bangladeshi studies rely on single universities and cross-sectional data, further weakening generalizability.

D. Limitations of the Studies

Overall, few studies rely on actual data analysis; many are reviews or perception-based surveys. Self-reported data, cross-sectional designs, and inconsistent definitions of activities limit findings. Meta-analyses offer scope but depend on weak primary studies [2]. Some use strong statistical methods [8], but many remain narrow in focus. Future research should employ longitudinal datasets, standardized definitions, and broader samples for more reliable results.

III. Methodology

A. Study Design and Participants

This study used a cross-sectional survey design through a Google Form to explore how student club participation, co-curricular and extra-curricular activities relate to academic performance and employment outcomes. The participants were university students of Bangladesh, including both undergraduates and recent graduates who were at least 18 years old and completed the survey voluntarily and anonymously.

B. Data Collection

For this research, we prioritize collecting primary data from university students. A detailed and well-structured survey questionnaire was developed, prioritizing the effects of club participation, co-curricular, and extracurricular activities, and how these factors influence or help them in their academic performance and employment. We collected responses from 163 students at different universities by sending the Google Form link directly to individuals, sharing it within related Facebook groups, and posting in the Reddit SurveyCircle community.

Our questionnaire consists of 20 well-written questions. We included the Bangla meaning of the questions and options for better understanding and readability.

Among the 20 questions, 18 are close-ended and 2 are open-ended. One open-ended question is to collect their university name, and the other is to gather the skills they have gained from clubs and extracurricular activities.

Among the 20 questions, some are simple Yes/No questions, where respondents indicate whether they agree or disagree with a statement. However, most of the questions are Likert scale questions, where respondents express their level of satisfaction, importance, or agreement with specific statements. The details of these questions are shown in Table II:

C. Data Pre-processing

Data pre-processing refers to how data is processed before being used by the researcher to get the aimed results. This is a very critical and important stage for data mining because the raw dataset contains several key issues such as data inconsistency, incomplete data, and noisy data. Working with these issues is very difficult for machines and could lead to misinterpretation.

For this research, at first, the variables of the actual questions in the Google sheet were renamed to facilitate multiple uses for result analysis. Additionally, some participant answers were manually edited as they fell under the same option category. Following this, null values were checked; since all questions were mandatory, there were no null values in the dataset.

Class imbalance occurs when one category dominates others, leading to biased machine learning results. In our dataset, the dependent variable 'Internship Job Offer' was imbalanced:

TABLE I

SURVEY QUESTIONS AND MAPPING OF VARIABLES

No.	Variable Name	Actual Questions
1.	Gender	What is your gender?
2.	Study Level	What is your current level of study?
3.	Department	What is your department or faculty?
4.	University	Which university are you currently studying at or did you graduate from?
5.	CGPA Range	What is your current CGPA range?
6.	Club Member	Are you currently a member of any student club or organization?
7.	Num Clubs	If yes, how many clubs or organizations are you a member of?
8.	Active Club Type	The club where you are most active - is it a departmental club or a central club?
9.	Activity Level	How active are you in your club?
10.	Club Position	Do you hold any role or position in the committee of the club?
11.	Extracurricular Involvement	Are you involved in any extra-curricular activities?
12.	Extracurricular Type	What types of extracurricular activities do you participate in?
13.	Extracurricular Hours	On average, how many hours per week do you spend on extracurricular activities?
14.	Skill Gained	Which skill have you gained the most from club and extra-curricular activities?
15.	Internship Job Offer	Has your involvement in club or extra-curricular activities helped you receive any internship; job offer or discount in any skill development training program?
16.	Industry Connection	Has your involvement in any club helped you build connections with industry professionals?
17.	Academic Effect	How have your club or extra-curricular activities affected your academic performance?
18.	Attendance Effect	During a semester, how often have your club or extra-curricular activities affected your class attendance?
19.	Job Prospect Help	How helpful do you think club activities will be for your future job prospects?
20.	Recommend Clubs	Would you recommend other students to participate in clubs or extracurricular activities?

TABLE II

DISTRIBUTION OF SURVEY QUESTIONS

Type of Question	Yes / No Question	Likert Scale Question
Number of Questions	6	14

112 'No' responses vs. 51 'Yes'. To address this, we applied SMOTE (Synthetic Minority Over-sampling Technique), which balances data by generating synthetic samples for the minority class instead of duplicating entries. It does this by creating new points between existing minority data and their nearest neighbours, improving algorithm learning and reducing bias. However, improper use may introduce noise or cause overfitting.

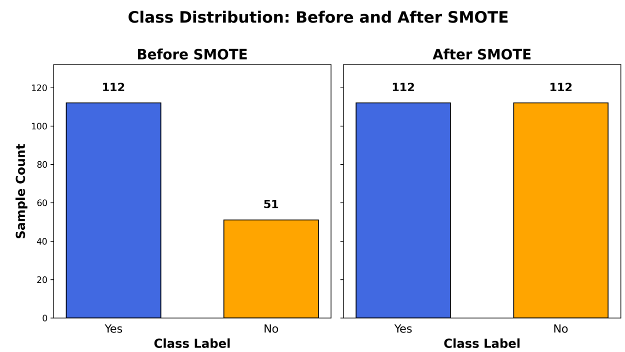

Fig. 1. Sample count before and after applying SMOTE technique

In our survey most of our data was categorical and on which we cannot directly apply machine learning algorithms. We got around this by applying label encoding, where we expressed categories as numeric values by giving each unique category an integer (e.g., 0, 1, 2). Using Scikit-learn's LabelEncoder, categorical values were converted to integers, making them suitable for algorithms expecting numeric input. However, label encoding may create an artificial ordinal relationship among categories and is therefore best applied to ordinal data (e.g., Low, Medium, High). For nominal categories without inherent order, alternative methods such as one-hot encoding are recommended.

Feature engineering is tool that improves model performance by creating, modifying, or selecting features to highlight useful patterns. In this research, we applied a chi-square test to identify features most strongly related to the dependent variable 'Internship Job Offer.' This allowed the dataset to emphasize the most relevant factors, helping the model learn more effectively and improve prediction accuracy.

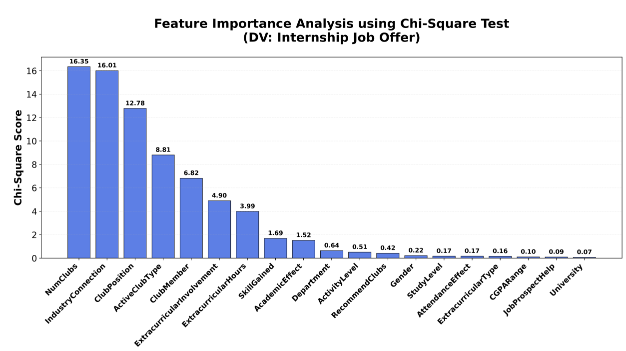

Fig. 2. Feature Importance Analysis using Chi-square

Normalization is a method to bring the data values in the same range, generally between 0 and 1. It is required because the different features of a dataset may have widely different scales (for example, age can be between 18–30, while income can be in thousands). If not scaled, the large values may overpower and confuse the model. Normalization makes all features contribute equally, which improves the performance and efficiency of machine learning algorithms. We applied z score and min max normalization.

Z-Score Normalization: The statistical formula for a value's z-score is calculated using the following formula:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Min-Max Normalization: The statistical formula for a value's min-max normalization is calculated using the following formula:

$$x_{norm} = \frac{x - \min}{\max - \min} \tag{2}$$

D. Model Training

After these pre-processing steps, the dataset was split for training, with a test size of 0.2, meaning 80% of the data was used to train the models and 20% for testing where random state was 0, which refers randomness is zero so that every time we run the code, we get the same result.

E. Applied Models

Decision Tree is a simple model that splits the dataset based on feature values to make decisions. It is used because it is easy to understand and interpret.

$$G(t) = 1 - \sum_{k=1}^{K} p_{t,k}^2 \tag{3}$$

Random Forest is an ensemble of many Decision Trees built on random samples. It is used to improve accuracy and reduce overfitting.

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_M(x)\} \tag{4}$$

K-Nearest Neighbors (KNN) is an instance-based, non-parametric learning approach known for deferring computations until classification. It assigns a class based on the majority of the nearest neighbors. It is used for tasks where similarity between data points matters.

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{5}$$

Naïve Bayes is a probabilistic model based on Bayes' theorem, assuming features are independent. It is used because it is fast and works well with categorical or text data.

$$P(C_k|x) = \frac{P(C_k)}{P(x)} \prod_{j=1}^{p} P(x_j|C_k) \tag{6}$$

CatBoost is a gradient boosting algorithm designed for categorical data. It is used because it reduces bias and avoids overfitting with ordered boosting.

Support Vector Machine (SVM) is a model that finds the best boundary (hyperplane) to separate classes. It is used for high-dimensional data and works well with complex decision boundaries.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i, \quad \text{s.t. } y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{7}$$

XGBoost is an efficient gradient boosting method that uses advanced optimization. It is used for its high accuracy, speed, and ability to handle large datasets.

Logistic Regression is a linear model for binary classification. It is used to estimate probabilities and serve as a strong baseline model.

$$\log \frac{P(y=1|x)}{P(y=0|x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{8}$$

IV. Results

Below is the correlation of each feature against the dependent variable 'Internship Job Offer', with positive correlations in green and negative in red.

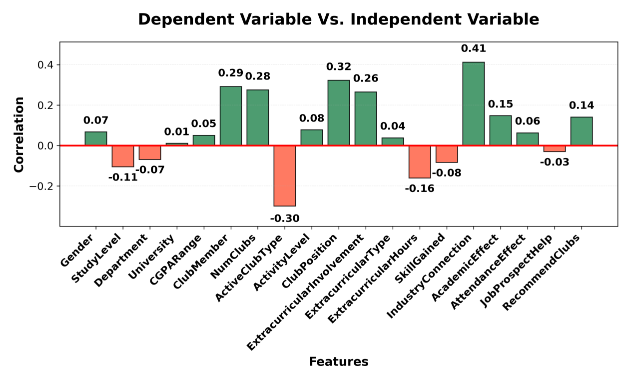

Fig. 3. Correlation with dependent variable

The correlation analysis reveals positive and negative correlations among student activities and internship outcomes. To the positive, Industry Connection ($r = 0.41$), Club Position ($r = 0.32$), Club Membership ($r = 0.29$), and Number of Clubs ($r = 0.28$) are the highest predictors, which suggest that the depth (leadership roles, industry connections) and breadth (membership in multiple clubs) of extracurricular activity make an extremely high difference in the likelihood of getting an internship offer. Conversely, there are some features that are negatively correlated. Active Club Type ($r = -0.30$), Extracurricular Hours ($r = -0.16$), and Study Level ($r = -0.11$) are all inversely correlated with internship performance. This indicates that high-time or loose-type participation without real engagement or skill mapping might not be beneficial to employability.

Varying strengths are noted in the contrastive examination of eight binary classification algorithms. Random Forest and CatBoost were optimal with the highest recall having 18 true positives and 3 false negatives each, although both had respectively 14 and 15 false positives that lowered precision.

Fig. 4. Confusion matrix of models

Naive Bayes was more balanced with 17 true positives, 4 false negatives, and 10 false positives. Decision Tree had the lowest false positive (8) and the highest true negatives (16) but slightly lower sensitivity with 16 true positives and 5 false negatives. Logistic Regression and KNN each attained 16 true positives but were affected by higher false positives (14 and 15). SVM and XGBoost each obtained balanced results, both with 16 true positives, mid-range false negatives (4 and 5), and false positives (12 and 13). Generally, Random Forest and CatBoost are the best performers for recall-bound applications, while Naive Bayes and Decision Tree are the best when false positive-minimization is the most important consideration. Considering both recall and precision trade-offs, Naive Bayes is the best balanced performer across all models.

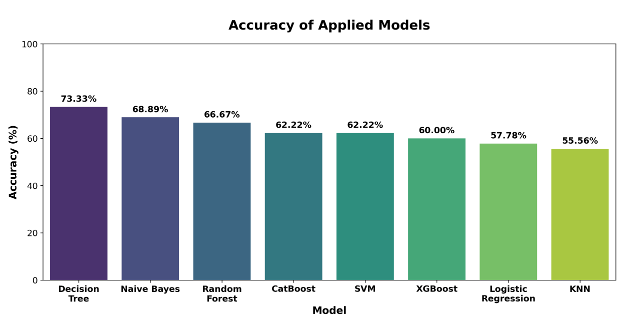

Fig. 5. Accuracy of the models

When comparing overall accuracy, Decision Tree tops the list at 71.11%, closely followed by Naïve Bayes at 68.89%. Random Forest, CatBoost, and SVM tie at 62.22%, while XGBoost (60.00%), Logistic Regression (57.78%), and KNN (55.56%) trail behind. These results suggest that tree-based learners, especially the single Decision Tree, capture the dominant patterns in this dataset most effectively, though ensemble models may offer more stability against overfitting.

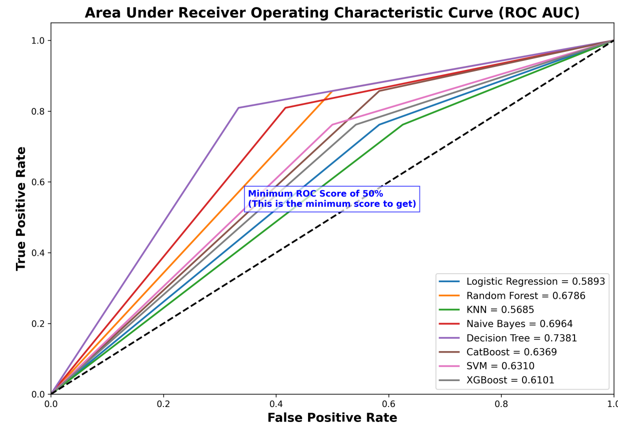

Fig. 6. Area Under the Curve (AUC) figure

The ROC curves corroborate these findings: Decision Tree achieves the highest AUC (0.7143), followed by Naïve Bayes (0.6964). Despite their lower accuracy, these models demonstrate superior ability to rank positives over negatives across thresholds. Random Forest (0.6369), CatBoost (0.6369), SVM (0.6310), XGBoost (0.6101), KNN (0.5685), and Logistic Regression (0.5893) exhibit moderate discriminative power, highlighting trade-offs between threshold-specific accuracy and overall ranking performance.

TABLE III
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.71111	0.66667	0.76190	0.71111
Naive Bayes	0.68889	0.62963	0.80952	0.70833
Random Forest	0.62222	0.56250	0.85714	0.67924
CatBoost	0.62222	0.56250	0.85714	0.67924
SVM	0.62222	0.57143	0.76190	0.65306
XGBoost	0.60000	0.55172	0.76190	0.64000
Logistic Regression	0.57778	0.53333	0.76190	0.62745
KNN	0.55556	0.51613	0.76190	0.61538

Decision Tree outperforms all other classifiers in terms of the accuracy measures show Decision Tree as the top one, highest in accuracy and F1-score and the most balanced in correctly classifying the positives and minimizing error. Naive Bayes follows, with highest recall, that means best at predicting positive cases but not quite so accurate. Random Forest and CatBoost also had high recall and low precision, while SVM and XGBoost were in the middle. Logistic Regression and KNN gave the worst results. In this scenario, precision measures the number of correctly predicted positives, recall measures the number of true positives picked up, and F1-score balances both. The Decision Tree is the most reliable model overall, but Naive Bayes is superior if picking up all positives is the biggest concern.

V. LIMITATIONS

This study also has some limitations that must be kept in mind while interpreting the results. Firstly, self-reported survey data could contain biases or errors because the respondents themselves may give incorrect answers or misreport and respond subjectively. Applying SMOTE and matching made the data balanced but relied on synthetic samples, which may not precisely show real patterns. The study also used only a single 80/20 train-test split, so the results might be varied using other splits. Since the study was conducted only on one university's students, the outcomes might not work for other universities or regions in Bangladesh. Requesting participants to answer all questions can lead to rushed or less thoughtful answers, which may affect the quality of data. Finally, distributing the survey mostly among familiar universities may create sampling bias. These points suggest the need for cautious interpretation and additional research to verify and amplify the findings.

VI. Conclusion

In short, our findings show that frequency of class attendance and study hours are the best predictors of academic achievement and internship success. Extensive participation in student organizations—measured in terms of membership, number of clubs to which a student belongs, and actual leadership roles—also hugely boosts opportunities for an internship. Predictive modeling confirmed the findings: a simple Decision Tree produced the best balance of accuracy and validity, while ensemble methods like Random Forest and CatBoost were most effective at predicting nearly all positive instances. Although oversampling and matching strategies helped create a fair comparison, the relative weak direct effect of living situation in residence supports that student achievement is far more dependent on engagement behaviours than it is on where students reside. These results suggest that universities must focus on courses that promote habitual attendance, institutionalized study habits, and meaningful extracurricular leadership in an effort to improve both grades and career prospects.

References

[1] S. R. Rahman, M. A. Islam, P. P. Akash, M. Parvin, N. N. Moon, and F. N. Nur, "Effects of co-curricular activities on students' academic performance by machine learning," *Current Research in Behavioral Sciences*, vol. 2, 2021.

[2] B. Shulruf, "Do extra-curricular activities in schools improve educational outcomes? A critical review and meta-analysis," *International Review of Education*, vol. 56, no. 5-6, 2010.

[3] K. R. Bartkus, B. Nemelka, M. Nemelka, and P. Gardner, "Clarifying the meaning of extracurricular activity: A literature review of definitions," *American Journal of Business Education*, vol. 5, no. 6, 2012.

[4] "A literature review of the impact of extracurricular activities participation on students' academic performance," *Journal Name*, vol. X, no. X, Year.

[5] J. S. Eccles, B. L. Barber, M. Stone, and J. Hunt, "The role of school-based extracurricular activities in adolescent development: A comprehensive review and future directions," *Journal of Social Issues*, vol. 59, no. 4, 2005.

[6] Iddrisu, O. Appiah, P. Appiahene, and F. Inusah, "A systematic review of the literature on machine learning application of determining the attributes influencing academic performance," *Decision Analytics Journal*, vol. 7, 2023.

[7] S. Bolton and J. Hattie, "The student perspective on employability," *Journal of Teaching and Learning for Graduate Employability*, vol. 5, no. 1, 2014.

[8] F. Islam, M. S. Islam, and M. M. Rahman, "Factors affecting the employability of private university graduates: An exploratory study on Bangladeshi employers," *Policy & Practice in Public Sector Reform*, vol. 1, no. 1, 2021.

[9] "The influence of learners' participation in school co-curricular activities on academic performance: Assessment of educators' perceptions," *Journal Name*, vol. X, no. X, Year.

[10] V. Dang, "Embedded and extra-curricular internships and activities for graduate employability," B.Sc. thesis, Aalto Univ., Mikkeli, Finland, 2023.

[11] M. M. I. Sabuj, R. K. Datta, and M. N. Rafiq, "The effect of extracurricular activities on the academic performance of university students: Evidence from Hajee Mohammad Danesh Science and Technology University (HSTU), Dinajpur, Bangladesh," *International Journal of Science and Business*, vol. 2, no. 3, 2018.

[12] M. M. Hossen, "The effect of extra-curricular activities on education of students of Daffodil International University," unpublished manuscript, Daffodil International University Institutional Repository, 2025.

[13] "Effect of extra-curricular activity on student's academic performance," *Journal of Armed Forces Medical College*, vol. 11, no. 2, 2015.

[14] "Overview of the students' perception of the impact of extracurricular activities on academic performance in Rajshahi University," *South Asian Journal of Social Sciences and Humanities*, vol. 6, no. 4, 2024.

[15] "The impact of extracurricular activities on learner's achievement in an EFL course at Daffodil International University, Bangladesh," OJER Publications, 2024.

[16] A. Begum, A. K. M. Asaduzzaman, M. H. K. Talukder, M. S. Hossan, and S. M. Akhter, "Extracurricular activities influencing academic performance of undergraduate medical students in Bangladesh," *Bangladesh Journal of Medical Education*, vol. 11, no. 1, 2025.