Machine Learning Approach to Predict SGPA and CGPA

Mohd. Saifuzzaman^{1*},Masuma Parvin¹, Israt Jahan¹, Nazmun Nessa Moon¹, Fernaz Narin Nur² and Syeda Farjana Shetu³

¹ Department of Computer Science & Engineering, Daffodil International University, Dhaka, Bangladesh

² Department of Computer Science & Engineering, Notre Dame University, Dhaka, Bangladesh

³Software Quality Assurance Department, BJIT Limited, Dhaka, Bangladesh

mohdsaifuzzaman.cse@gmail.com,parvin.ged@daffodilvarsity.edu.bd,{isratjahan.cse,moon}@diu.edu.bd, fernazcse@ndub.edu.bd,

syeda.farjana@bjitgroup.com

Abstract— The prediction of SGPA and CGPA is beneficial to university students. Students will easily get an estimate of their final outcome from this project. As a result, the students will be able to brace themselves for a successful outcome. Students pass the day by participating in a variety of events. Students use social media sites such as Facebook, Instagram, and Twitter. They engage in various hobbies such as playing mobile games, listening to music, among others. As a result, they were able to move several times with these tasks. As a result, if a student spends so much time doing any of those things, she will not be able to achieve a successful grade because of the experiment; students can develop a research routine or guideline that they can apply to their other tasks. Additionally, students' behaviors will forecast their outcomes. The Authors will now see machine learning in Python being used all over the place. After that, The Authors created a smart SGPA and CGPA prediction project, as well as the results on students. The findings are predicted using the Nave Byes algorithm. The Nave Byes algorithm is a simple but effective prediction algorithm. It is a machine learning algorithm as well. As a result, students will be given an estimate of their final exam scores. They can prepare them to make a good result by following the routine of the SGPA & CGPA prediction project.

Keywords—CGPA, Machine Learning, Prediction and SGPA

I. INTRODUCTION

Nowadays, the world goes so faster. This is the era of modernization & competition. In the education system there is too much competition between students. Basically, in the universities there is the educational system is different from school or college. In universities there has some curriculums or activities for students. In the universities there has, attendance mark, quiz marks, assignment marks, presentation marks, project marks, mid-term marks, & final exam marks. So, if the students achieve the good marks in all segments then they have to study more and sincerely. Students pass their time by doing all other activities. That's are also important for a human but if students pass their all times or more times using social media or something like that, then they will not be able to make a good result. That's why The Authors built a project that is called, Prediction of The Final Exam SGPA & CGPA. In the project, The Authors tried to make a good routine for students. As a result, students can follow the routine and make a good preparation for exam to get a good result. Also, The Authors collected data from different kinds of university students. There are some students who got good marks in all the segments and there also those students who did not get the enough good marks and also there are some students who got worst marks in all the segments. If a student got less mark than his/her expectations but he/she wants to make a good result then he/she will use the project for the prediction SGPA. He/she will get suggestion from the project that how he/she would take preparation for a good result. In the project The Authors used machine learning/ Deep learning.

II. LITERATURE REVIEW

In a research done by Lubna Mahmoud Abu Zohair [1] who made a project Prediction of Student's performance by modelling small dataset size. Prediction of student's performance is an urgent desire at this time. So, in this situation they think to help at-risk students and assure their retention, providing the excellent learning resources and improving the university's ranking and reputation. But that was not so easy to be achieved for startup to mid-sized universities, especially those which are specialized in graduate and post graduate programs, and have small students' records for analysis the students result. Also, in most researches that were aimed to classify or predict, researchers used to spend much efforts just to extract the important indicators that could be more useful in constructing reasonable accurate predictive models. And the finally they will look at the selected features while training the dataset on different machine learning algorithms, like in (Comendador, Rabago, & Tanguilig, 2016; Mueen, Zafar, & Manzoor, 2016). Navin venkat (Indian Institute Of Science) & Sahaj Srivastava (Birla Institute of Technology & Science)[2] are made a project on Predicting Student Grades using Machine Learning. They predict students grade point with machine learning. Also the used data classifier as data mining. They model the prediction of final Grade as a classification problem. Each grade is assumed to be a class. Hence, they have 9 classes (A through E, and NC), since those denote the students who took the course till completion. For these tests they ignore the class withdrawn (W) because most of the data is missing for such records. Since the dataset is very small (197 records for scores, after removal of they as mentioned above), they perform a stratified 5-fold cross-validation for each of the classifiers to observe and compare their stability. The stratification aids in handling the class imbalance problem. They have used Decision Tree, Naive Bayes, SVM (with linear, RBF and sigmoid kernels).

Khurum Nazir Junejo and Eman Eman [3] both are made a project on Grade Prediction Using Supervised Machine Learning Techniques. They have made a first attempt at predicting the final grade of students in undergraduate courses based on their performance in the same course prior to the final exam and the type of the course and the teacher teaching that course. After applying various data preprocessing and machine learning algorithm on 2500

course records, they have achieved quite a remarkable prediction performance (varying between 73-96 % accuracy). They have the following useful conclusions: Grade of more than 96% students can be accurately predicted even without the result of the final exam. This implies that the performance of students in various instruments before the final exam is highly correlated to their performance in the final exam. Another interesting point is that the F grade is the most easiest to predict followed by C grade, while B+ is the most difficult grade to predict. After all the number of students whose grade slip because of their performance on the final exam is more than those who improve it because of final exam and finally rule induction classifier outperforms ID3 classifier on student performance prediction. Moon et al. [6] suggested a natural language processing method that can detect video. Natural language processing has recently gained great attention as a novel interdisciplinary field. Salehin et al. [7] suggested artificial intelligence and LSTM strategies for a rainfall forecast model that is readily available for determination. For this method of the implementation process, the deep learning approach is most important and its accuracy is established. They used 6 parameters in their paper. 76% accuracy was achieved by analyzing all the data. Salehin et al. [8] suggested a model RHMCD that helps to achieve the desired objective using machine learning algorithms. The algorithms that were experimented with are: Naive Bayes classifiers, logistic regression, and Support Vector Machine The sentiment analysis technique was used for gaining the report of mental conditions. The decision tree approach was used to measuring the level of depression. Salehin et al. [9] predicted the depression level of excessive use of the mobile phone. Two machine learning algorithm decision trees and the Linear Regression algorithm are used to detect depression. Salehin et al. [10] have designed agriculture technologies. A large loss of agricultural products is caused by various forms of influenza, fungal and bacterial infections. In this article they use the algorithm Scale-Invariant Transform Feature (SIFT) to classify crop conditions based on different datasets. Finally, the solution was provided using SMS services and live web portals. Talha et al. [11] point out the big negative side issue and its various causes such as emotional imbalance, depression, stress, loneliness, etc. The data was collected using three methods physical, virtual, and medical reports. The 71% optimistic theorem of Naive Bayes shows the negative effect of human behavior. In SVM they set an adverse and positive parameter for measuring. Finally, they equate the effects of our proposed specialization with the results of the three basic points of reference. Shetu et al. [12] in this research is intended to create an algorithm to recognize the Bangladesh style, whether it is from a given Bangladesh paragraph input in Sadhu Bhasha or Cholito Bhasha. It is a contribution to discovering the Goruchondali Dosh, a common grammatical error in written language Bangladesh, which has shown that there are a lot of investigative works to detect errors in Bangladesh grammar while other linguistics researchers have a common pattern. Shetu et al. [13] predicted the performance of the students through educational data mining. The whole prediction was depended on the overall academic status and environment of the place they are in. Hosen et al. [14] proposed a way to play music and to alter

the screen wallpaper based on emotional recognition by facial expressions in live video streams. To extract different facial features, they used a webcam to build their own dataset according to facial expression and Haar classifiers are used here. Then we produce an emotional and user interest-based playlist and also create a model of the shifting wallpaper so that we play songs that suit our moods and turn our desktop wallpaper into happier images if we are saddened. Ullah et al. [15] proposed the Adjacent Collaborative MAC Protocol (ACDM) for wireless networks, which is a modern directional MAC protocol. The aim is to increase the efficiency of the drive and delay and to reduce the overall wireless network. Sarker et al. [16] classified the practitioner's behavior in medical information. They used datamining to complete the research. Moon et al. [17] used the Bangla 'Banjonborno' and 'Sorborno' characters for the recognition system. In the learning stage, the characters are scanned and converted into m x n matrix which is reduced to 16 X 16 using the Matlab resume scaling function. The function matrix is then transmitted to the multilayer Neural Network. The supervised method of learning is used to monitor the feedforward algorithm. This recognized system can recognize any size of character with an accuracy above 74%

III. METHODOLOGY

Research methodology used to identify select, process and analyze information. Now in this chapter The Authors will discuss & describe about the research methodology. Moreover, tools for the research project, data collection, research topic, pre-processing, processing, statistical analysis, and its implementation will be discussed in this session.

The steps of research methodology is given below at fig.1.



Fig.1. Research Methodology at a glance

A. Data Collection Procedure

In survey, The Authors used a questionnaire as dataset. These data are mainly selected according to the effects of mobile use on a student's academic result, health condition, social relationship, and psychological status

• Data Pre-processing

Data pre-processing is a processing that means to the pre phase of processing datasets. Generally raw data sets are not able to perform operations and generate expected outcome. As a result, data pre-processing is required. Also it is considered to be one of the most important parts of research. In this phase, The Authors have collected more than 500 surveys. Here, to predict the final exam result.

• Data Organizing

Data organizing is a system of organize data. So, for organizing data, The Authors have tested and trained the data and saved them in two folders. The Authors have also

used validation folder to check train data validation in data organizing.

Labeling Data

At first The Authors have to do label the data. It's called data labeling. The Authors labeled the dataset with label encoder with python. Fig.2. shows the data labeling by Data Encoder.



Fig.2. Data Labeling by Data Encoder After labeling data The Authors got numeric dataset. Given below with Fig.3.



Fig.3. Numeric Dataset

• Data Storing

Where data can be stored it's called data store So, The Authors have to store data. Here The Authors store all the data in Google drive because it makes the work easier. In Google drive, The Authors can use those online stored data in the project The Authors haves saved all the data as CSV files. The advantage of google drive that data cannot be lost.

Statistical Analysis

Table 3.1 the trained data amount 504:

Current Semester	Student Amount
1,2	31,39
3,4	41,43
5,6	45,40
7,8	38,43
9,10	48,42
11,12	41,53

IV. EXPERIMENTAL RESULTS & ANALYSIS

From the collected data The Authors have run the dataset to create a model. It provided us with the desired the output. Then The Authors created each features in dataset.

A. Experimental Setup

In experimental setup for our model and code implementation we have collected the data first. The procedure is given below:

 We are working with prediction of impacts on social media using, paly games, and leisure time

- spend, study time, attendance marks, assignment marks, quiz marks, presentation marks, midterm marks etc. So, we had to collect student data on these topics.
- Taking survey we have spent more time. After all in this pandemic situation it is so tough to collect data from students in physically.
- We collected data from online by using Google Form.
- After labelling data, they become perfect for further use.
- After that, we have converted data to numeric type.
- After that, we took the normalized and finalized data so that we can begin the training.
- Then we have preprocessed our data.

B. Correlations for each Features

Here The Authors can see in the fig.4. here is all features are correlated. From the beginning The Authors got, quiz marks, mid-term marks, assignment marks, presentations marks and then from these marks The Authors can predict a SGPA for the students. But here also some activities those will effect in the result. Those are, academic study, spend time on mobile, spend time on online, spend time on mobile, spend time on social media, spend time for online study, spend leisure time. The result is dependent on those activities also.

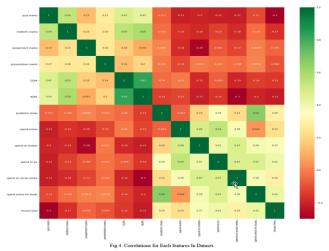


Fig.4. Correlation of Each Feature in Dataset

C. Effect of Activities

There are so many activities in student daily life. They do study, time pass with social media, video game playing, leisure time, online study etc. All these activities have to a routine. That's why The Authors created this project for help the student make their routine and target a good SGPA or CGPA.

D. Assignment

Assignment is a very important task of all subjects. It is mandatory to submit for make a good result. It carries 5 marks out of 100. So, it's not so neglected topic. Teachers give a topic to the students and give instructions for write an assignment. After that students do their assignment work carefully? In an assignment there would be some important topics, some important objectives, case study about the main

topic etc. If students do this assignment well, then it is helpful to get a good result. The graph is showing at Fig.5.



Fig.5. Effect of assignment

E. Quiz

In a semester there are another important test. That is quiz. Quiz is a plus point to make a good SGPA or CGPA. If students do well in the quiz, then they will be able to make good SGPA. Fig.6. shows that.



Fig.6. Effect of quiz

So, here is the graph of quiz effect.

F. Presentation

Presentation is a good way to make a well communication skill. Students who can make presentation and able to give presentation in front of audience, they are very confident. Presentation creates confidence level up. Every student must have this skill. Presentation segment carries 7 marks out of 100 for each subjects or courses. Here is the graph for effect of presentation. Fig.7. shows that.

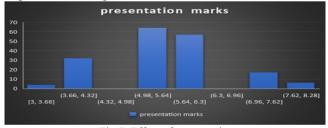


Fig.7. Effect of presentation

G. Mid Term Exam

Mid-term exam is a big part of get marks out of 100. In the mid-term there will be 25 marks on each subjects. So, if students get good marks in the mid-term exam also, so it is very easy for those students to make a good sgpa and cgpa. In the project The Authors suggest students how to make a good result in the final exam. Students will get a good suggestion from the project. Here is the effect of mid-term exam graph at Fig.8.



Fig.8. Effect of mid-term exam

So, Authors can see from the graph, the effect of mid-term exam. Who will get good marks in the mid-term they will be able to make a good result in the final exam easily.

H. Effect of Semester

Here, at the Fig.9. Authors can see from the graph, here is the effect of semester on results. The graph shows different types of results with semester.



Fig.9. Effect of semester

I. Effect of SGPA

After all these The Authors got the effect of SGPA graph. At Fig.10. the Authors can see this graph:



Fig.10. Effect of SGPA

So, this is the graph for effect of SGPA.

J. Effect of CGPA

Here The Authors got, both SGPA & CGPA from this graph. The graph is given below at Fig.11.



Fig.11. Effect of CGPA

V. ALGORITHMS USED

In this research, Authors need best accuracy. The Authors have run different types of algorithm to know which one give a better output. That's why, at WEKA The Authors have run linear regression model and logistic regression algorithm at python.

A. Linear Regression

At WEKA The Authors used linear regression. Linear regression model show the CGPA and SGPA accuracy 87%. So, The Authors got 87% accuracy on SGPA and CGPA. The following fig.12. & 13 Shows in the bellow:

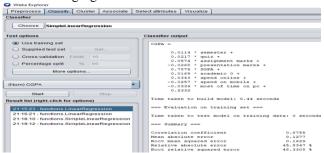


Fig.12. Linear regression (CGPA)

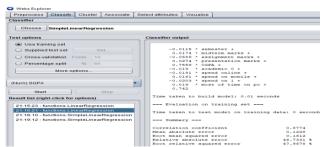
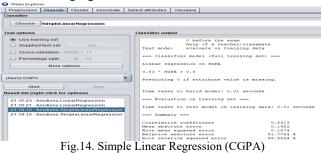


Fig.13. Linear regression (SGPA)

B. Simple Linear Regression:

Here, The Authors used simple linear regression and The Authors got 83% accuracy on CGPA and SGPA.

The following fig.14. & 15 Shows in the bellow:



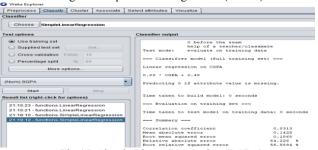


Fig.15. Simple Linear Regression (SGPA)

C. Random Forest (CGPA & SGPA)

In classification on WEKA The Authors use tree and choose Random Forest tree and show the correlation coefficient for CGPA rate is 98.34%. The following fig.16 shows bellow:



Fig. 16. Random Forest (CGPA)

In classification on WEKA The Authors use tree and choose Random Forest tree and show the correlation coefficient for SGPA rate is 98.63%. The following fig.17. shows bellow:



Fig.17. Random Forest (SGPA)

D. Random Tree (CGPA & SGPA):

The Authors take another tree and choose Random tree and show the correlation coefficient for CGPA rate is 99.99%. The following fig.18. shows in bellow:

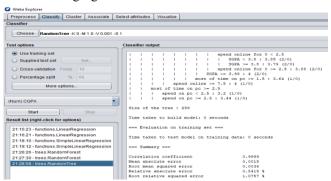


Fig.18. Random Tree (CGPA)

The Authors take another tree and choose Random tree and show the correlation coefficient for SGPA rate is 1. The following fig.19. shows in bellow:

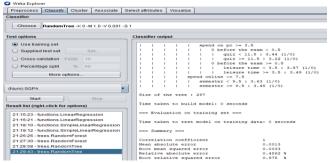


Fig.19. Random Tree (SGPA)

E. Decision Table (CGPA & SGPA)

The Authors take another tree and choose Decision Table and show the same correlation coefficient rate for CGPA. The rate is 81.82%. The following fig.20. shows in bellow:

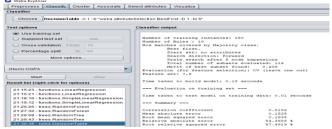


Fig.20. Decision Table CGPA

The Authors take another tree and choose Decision Table and show the same correlation coefficient rate for SGPA. The rate is 84.82%. The following fig.21. shows in below:

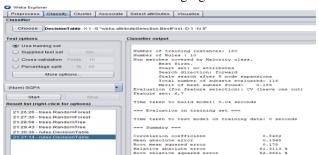


Fig.21. Decision Table (SGPA)

F. Logistic Regression using Visualizing Confusion Matrix

Use a confusion matrix to visualize the results of the model in the form of a confusion matrix using numpy, seaborn and matplotlib. Here, fig.22. shows the matrix value

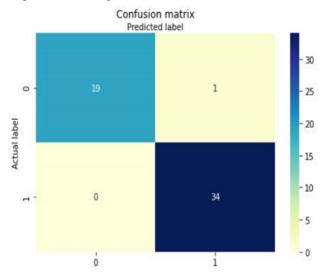


Fig.22. Confusion Matrix

G. Effect of Result

From all those researches The Authors can say that, the effect of result will be depend how student manage their time. That means, how many times they spend on which activities. If students give more time to study, then they will get good marks and able to be make a good result. But if students give more time to other activities like social media, video gaming, leisure time, and then they are not able to get a good mark. As a result, they will not be able to make a good result. So, it's very important to follow the routine for the students for make a good result.

VI. CONCLUSION

In this modern era technology is so advanced e.g. IoT [18-21], Machine Learning [22] and so on plays a vital role where in our study we worked on Machine Learning approach to predict our desire result to contribute the research area. The Authors all are probably technology dependent in the daily life. But it has some advantages & disadvantages at the same time. Controlling big data is a complex problem in the technological field. Nowadays, the world goes so faster. This is the era of modernization & competition. In the education system there is too much competition between students. Basically, in the universities there is the educational system is different from school or college. In universities there has some curriculums or activities for students. Prediction of the SGPA or CGPA is

extremely helpful project for students. The Authors created this project using machine learning.

References

- Lubna Mahmoud Abu Zohair- "Prediction of Student's performance by modelling small dataset size", 2019. [1]
- Machine Learning", Nov 2018.

 Khurum Nazir Junejo & Eman Eman— "Grade Prediction Using Supervised Machine Learning Techniques". 2014.

 Prepare organizational data in Workplace Applying [2]
- Prepare organizational ..docs.microsoft.com How to Write a Res data in Workplace Analytics
- Research Methodology in How Four Steps -Scribbrwww.scribbr.com
- Moon, N. N., Salehin, I., Parvin, M., Hasan, M. M., Talha, I. M., Debnath, S. C., Nur, F. N., and Saifuzzaman, M. A natural language processing based advanced method of unnecessary video detection and the framework model In International Journal of Electrical and Computer Engineering (IJECE), Indonesia.

 Salehin, I., Talha, I. M., Hasan, M. M., Dip, S. T., Saifuzzaman, M., &

- Salehin, I., Talha, I. M., Hasan, M. M., Dip, S. T., Saifuzzaman, M., & Moon, N. N. (2020, December). An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network. In 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 5-8). IEEE. Salehin, I., Dip, S. T., Talha, I. M., Rayhan, I., Nammi, K. F. "Impact on Human Mental Behavior after Pass through a Long Time Home Quarantine Using Machine Learning", International Journal of Education and Management Engineering (IJEME), Vol.11, No.1, pp. 41-50, 2021. DOI: 10.5815/ijeme.2021.01.05
 Salehin, I., Talha, I. M., Moon, N. N., Saifuzzaman, M., Nur, F. N. & Akter, M. "Predicting the Depression Level of Excessive Use of Mobile Phone: Decision Tree and Linear Regression Algorithm" on 2nd International Conference on Sustainable Engineering and Creative Computing (ICSECC-2020), 16 17 December 2020, President University, Indonesia. Indexing: IEEE Xplore, EI-Compendex, SCOPUS. University, SCOPUS.
- Salehin, I., Talha, I. M., Saifuzzaman, M., Moon, N. N., & Nur, F. N. Salehin, I., Talha, I. M., Saifuzzaman, M., Moon, N. N., & Nur, F. N. (2020, October). An Advanced Method of Treating Agricultural Crops Using Image Processing Algorithms and Image Data Processing Systems. In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (pp. 720-724). IEEE. Talha, I. M., Salehin, I., Debnath, S. C., Saifuzzaman, M., Moon, M. N. N., & Nur, F. N. (2020, July). Human Behaviour Impact to Use of Smartphones with the Python Implementation Using Naive Bayesian. In 2020 11th International Conference on Computing, Communication and
- 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

 Shetu, S. F., Saifuzzaman, M., Parvin, M., Moon, N. N., Yousuf, R., & Sultana, S. (2020, July). Identifying the Writing Style of Bangla Language Using Natural Language Processing. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

 Shetu, S. F., Saifuzzaman, M., Sultana, S., Yousuf, R., & Moon, N. N. (2020). Students performance prediction through education data mining.
- (2020). Students performance prediction through education data mining depending on overall academic status and environment. In 3rd International Conference on Innovative Computing and Communication
- International Conference on Innovative Computing and Communication (ICICC-2020) on.

 [14] Hosen, P., Himel, N., Adil, M. A., Moon, M. N. N., & Nur, M. F. N. (2019). Music Playing and Wallpaper Changing System Based on Emotion from Facial Expression. In Emerging Technologies in Data Mining and Information Security (pp. 79-87). Springer, Singapore.

 [15] Ullah, M. S., Nur, F. N., & Moon, N. N. (2015). Optimization of wireless ad-hoc networks using an adjacent collaborative directional MAC (ACDM) protocol. International Journal of Computer Applications, 114(3).

 [16] Sarker, K. U., Moon, N. N., & Ahmed, S. (2011). Classifying the practitioner's behavior in medical informatics by using data mining. International Journal of Computer and Information Technology, 1(2), 23-25.
- 23-25.
- [17] Moon, N. N., & Sarker, K. U. (2011). Size Independent Bangla
- Moon, N. N., & Sarker, K. U. (2011). Size independent banga Character Recognition System by Multilayer Feedforward Neural Network IJCIT, 2(1).

 Moon, N. N., Nur, F. N., Saifuzzaman, M., Sharmin, F., Hosen, P., & Shetu, S. F. (2020). An Efficient Development of Automated Attendance Management System. 3rd International Conference on Innovative Computing & Communications (ICICC) 2020. Attendance Management System. 3rd Internatio Innovative Computing & Communications https://doi.org/10.2139/ssrn.3563387.

- https://doi.org/10.2139/ssrn.3563387.
 Mohd. Saifuzzaman, Ashraf Hossain Khan, Nazmun Nessa Moon and Fernaz Narin Nur "Smart Security for an Organization Based on IoT" on International Journal of Computer Applications (IJCA), Volume 165, Number 10, May 2017, New York, USA.
 Mohd. Saifuzzaman, Nazmun Nessa Moon & Fernaz Narin Nur, "IoT Based Street Lighting and Traffic Management System" on IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2018, BUET, Dhaka, Bangladesh.
 Mohd. Saifuzzaman, Syeda Farjana Shetu, Nazmun Nessa Moon, Fernaz Narin Nur and Mohammad Hanif Ali "IoT Based Street Lighting Using Dual Axis Solar Tracker and Effective Traffic Management System Using Deep Learning: Bangladesh Context" on 11th
- Using Dual Axis Solar Tracker and Effective Traffic Management System Using Deep Learning: Bangladesh Context" on 11th International Conference on Computing, Communication And Networking Technologies (ICCCNT 2020), July 1-3, 2020, IIT Kharagpur, West Bengal, India.

 Saif Mahmud Khan Durjoy, Abu Mohammed Rabbani Rafi, Zerin Nasrin Tumpa and Mohd. Saifuzzaman, "A Comparative Study on Prediction of Dengue Fever Using Machine Learning Algorithm" on The First International Conference on Advances in Distributed Computing and Machine Learning(ICADCML-2020) on 30th and 31st January 2020.