May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

Evaluation of Faculty Teaching Performance Using Students' Heart Rate Data

[1]Mu Lin Wong, [2]Senthil, S.

[1]School of Computing and Information Technology, REVA University, Bengaluru, India
[1]joshuawml@yahoo.com
[2]School of Computer Science and Applications, REVA University, Bengaluru, India
[2]dir.csa@reva.edu.in

Article Info
Volume 83
Page Number: 3753-3759
Publication Issue:
May-June 2020

Abstract
Existing teaching performance evaluation depends on data collected through questionnaire and class observation. Both are intrinsically polluted by bias factors. Therefore, questionnaire requires high volume while class observation requires expert input, to reduce bias. This study experiments using heart rate data of 32 students in 3 subjects over 30 classes each to derive many attributes that can be used to evaluate teaching performance. The empirical evidence showed that Peak attribute is the most robust attribute in relating to students' scores, having the highest Pearson's Correlation Coefficient with 0.01 significant level, followed by Up, Down and Low. These attributes can be used on the first class of the subject as they don't fluctuate much. Students' rating on teachers is found to be inaccurate in evaluating teachers' performance. A good teacher is one who can stimulate students in class, resulting in high maximum heart rate, high class engagement and maximum tiredness after class.

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Keywords: *Educational Data Mining; Students' Evaluation of Teaching; Heart Rate Fluctuation; Pearson's Correlation; Teaching Performance Evaluation.*

1. Introduction

Traditionally, the evaluation of teaching performance depends mostly on Likert-scale questionnaire-based analysis. The bias resulted in the analysis is often traced to the way data is collected, either having the teachers being evaluated collecting the questionnaires or doing the questionnaires collection before the teachers submit the results of the students. At other times, students compromise in filling in the questionnaires to appease the higher authorities.

Students' engagement during a class or a lecture can be used as an added feature in building and testing a classification model. It is a real time and dynamic data. Therefore, the more of this data being collected over time, it will contribute to the accuracy increment of the model being created base on this data. Besides being used as a feature in the classification model, it can be used to identify which lecture a particular student or a group of students isn't engaging and the teacher can make efforts in revising the topic to an individual student, a selected group of students, or the whole class. The teacher can also alter his or her teaching pedagogy in order to make student learning effective.

Another method developed to counter the weaknesses of questionnaire-based teacher analysis is class observation. Unfortunately, this empirical study is also subjected to human errors and bias-prone analysis. An advanced version is using audio and video recording to gauge the engagement level of the students and the teaching effectiveness of the teachers. Since the equipment applied in this kind of research is expensive and the analysis of the data required high expertise, the advancement of such study is often hampered.

The authors of this study are looking for a more cost-effective and non-bias way of data collection, in

May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

order to analyze the efficiency and effectiveness of the teaching faculty during their theory classes. The proliferation of health trackers that collect physiological data and the free cloud storage are making the data collection and storage cost-effective. Therefore, the aim of the authors in this study is to ascertain the authenticity of using heart rate data of first year students as an efficient and effective way to gauge the performance of the teaching faculty in a university setting. The derived attributes of the heart rate data, the rating of each faculty member collected privately from the students and the examination results of the students are used in Pearson's Correlation Test to perform this analysis.

The objectives of this study can be arranged as below:

To determine which derived heart rate attribute(s) can be used to measure the performance of the teachers.

To compare the results of students' feedback and the actual exam results by the students.

To ascertain when is the earliest for the cumulative derived heart rate attribute(s) chosen in objective 1 to stabilize.

This paper is arranged in the following order. The Literature Review segment presents the survey results of literature related to evaluating teachers' performance while the Materials & Methodology segment describes the methodology carried out to analyze derived heart rate data of students to determine the attribute(s) which can be used to evaluate teachers' performance. The following segment discusses the results and its implication. Finally, the Conclusion segment provides a general conclusion and a mention of the scope of future research.

2. Literature Review

In a review paper written by Romero and Ventura in 2010 [1], Romero & Ventura articulated that one of the many uses of Educational Data Mining is to provide feedback to support course teachers to improve students' learning and to provide remedial help whenever necessary. Many studies have been made since then focusing on analyzing and predicting the academic performance of students based on various student attributes, as mentioned in another review paper [2]. However, there was no mention of analyzing the performance of teachers using data mining methods until a study was done in 2016 [3]. The study used data obtained from a questionnaire given to 2850 students and C5.0 appears to be the best classifier with more than 90% accuracy. The analysis of the attributes demonstrated that the subject and the students' interest are more important than the teacher's behavior during a teaching performance evaluation.

In a recent study [4] in Pakistan, an aspect based sentiment analysis on 5000 students' textual feedback using deep learning was performed to evaluate faculty teaching performance. It was found to be above 90% in accuracy was achieved using modified Long Short-Term Memory (LSTM) model. In another similar study [5] conducted in Taiwan for 20,000 students using attention LSTM classifier, recorded an accuracy above 90% when time series factor is included in text sentiment analysis. Another study [6] presented the feasibility of teaching performance evaluation using data collected in a smart campus environment and then applying three algorithms to select the best attributes to measure teacher efficacy, eliminating the erroneous method of summing up all index scores.

Student survey questionnaire is a common way to collect data to evaluate course content [7] or teaching faculty [3]-[5]. However, there is a risk that students' answers on a questionnaire can be biased. Moreover, the practice of using questionnaires and class observation has been the most common in teaching and learning environments. A study [12] done on 39 principals in Israel showed that principals tend to inflate the evaluation on teachers in order to maintain a good relationship with teachers, among other reasons. Another study [13] based on the evaluation of 85 undergraduate students on teachers revealed that though the students were able to pinpoint the correctness of the subject taught, their evaluations were marred by gender bias and lack of understanding of educational dimensions.

Feistauer & Richter [14] cautioned that results from students' evaluation of teaching should be interpreted with caution due to the empirical evidence of bias found in the study. They reasoned that the teacher evaluation seemed more like a reflection of the likability of a teacher rather his/her teaching ability. Thielsch et al. [15] argued that biases due to non-response in students' evaluation of teaching can be reduced or eliminated by maintaining proper communication with the students before and during such evaluations.

Fauth et al. [8] empirically proven that a teacher's pedagogical content knowledge, self-efficacy and teaching enthusiasm positively correlate to students' interest during class. Bradford & Braaten [9] highlighted that teacher evaluation can demoralize teachers should the assessment focused too much on management and vertical accountability. Another study [10] also indicated that student evaluation on teaching affect teachers' self-efficacy, whether the feedback was reliable or not. Therefore, there is a tendency among teachers to not appreciate feedback unless the feedback is specific, frequent and evidence-based [11].

There is a new dimension opening up to evaluate teaching performance specifically and frequently. It's the use of physiological health trackers to collect the physiological data of students to evaluate their class engagement, which can be used to gauge the efficacy of teachers in class. With the proliferation of reliable health trackers available in the market, the data collection can be obtained easily. This literature review motivated us to proceed with the data collection and analysis discussed in the next segment.

May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

3. Materials & Methodology

Datasets

A class of 32 first year university students aged between 18 and 20 volunteered in the heart rate data collection during class lectures. C Programming, Digital Electronics and Mathematics were the three subjects chosen where the students' heart rate will be recorded. Each class last for approximately 1 hour. Thirty class hours for each subject were recorded. The data were then derived into 13 attributes for each student, as shown in Table I.

Table 1: Attributes of the dataset and their definitions

SI No.	Attributes	Definition
1	Name	Student name
2	Max	Maximum heart rate in 1 hour class time
3	Min	Minimum heart rate in 1 hour class time
4	Range	Max minus Min
5	Engage	Percentage of class engagement ((Ave-Rest)/(Stim-Rest))
6	Time	Number of seconds in attending class
7	Stim	Average heart rate when stimulated in playing mobile game
8	Rest	Average heart rate when resting with eyes closed
9	Slope	Slope of the regression line of the 1 hour heart rate graph
10	Attend	Percentage of class attendance
11	Up	Sudden surge of 10BPM or more within 10 seconds
12	Down	Sudden drop of 10BPM or more within 10 seconds
13	Peak	Upward spike with amplitude of 10BPM or more within 30 seconds
14	Low	Downward spike with amplitude of 10BPM or more within 30 seconds
15	SGPA	Semester Grade Point Average

Student name and SGPA are other attributes not derived from heart rate data. Stimulated heart rate is the average of 10 weekly 3-minute heart rate data collected when a student is playing mobile game. Resting heart rate is the average of 10 weekly 3-minute heart rate data collected when a student is resting with eyes closed. Engage is calculated by taking the average heart rate of a student within an hour class and minus the resting heart rate, then divide by the difference between stimulated and resting heart rate of that student. The SGPA is rated between 0 (failed) and 10 (outstanding). The SGPA value of 5 means pass, 5.5 means average, 6 means above average, 7 means good, 8 means very good and 9 means excellent. Other attributes are straight forward and easily comprehensible.

For each subject, a dataset of the above attributes are prepared and then tested using the Pearson's Correlation Test available in SPSS software tool, to ascertain which two attributes are correlated and how much they are related.

Pearson's Correlation Test

This is a statistical test that is based on the method of covariance to verify if two independent variables are positively related, negatively related or not related at all. The result of the test is the Pearson's Correlation Coefficient value, also known as Pearson's R value. The r values range from -1 to 1. -1 denotes the two attributes are absolutely inversely related to one another while 1 denotes that the two attributes are perfectly correlating to one another. 0 denotes that there are no relationships between the two attributes. Any value above 0.5 denotes a moderate to strong positive correlation between two attributes while any value below -0.5 denotes a moderate to strong negative correlation between two attributes. In this study, our focus is on attributes with correlation value that is above 0.5 or below -0.5.

Another important consideration while interpreting the Pearson's R value is the significance level. For example, if the significance level is 0.05 or 5%, it means that the risk of concluding that two attributes are correlated while there actually is no correlation is 5%. In this study, we focus on significance level of 0.05 and 0.01. Of course, the smaller the significance level indicates the less of risk in making wrong conclusion.

Experimental Procedures

The heart rate data of 32 students were collected and preprocessed according to the various attributes mentioned in Table 1. Then, the average of the thirty 1-hour class of each subject is calculated before applying Pearson's Correlation Test. The attribute pair having R value of more than 0.5 (or less than -0.5) and significance level of 0.05 or 0.01 will be identified and discussed. The attribute pair consistent in each subject should be the most critical attribute in determining the efficacy of the teacher in classroom teaching.

Then the average values of each attribute for all 32 students in each subject is calculated to observe if there is a relationship between the SGPA attribute with each other attribute. We can observe which subject is having a higher academic score and consider whether the score is directly or inversely related to other attributes. This should help us to discover critical patterns to evaluate teachers.

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

Based on the most prevalent attributes identified earlier, the cumulative values over the 30-class interval for each subject is kept on a line graph to estimate when the earliest the graph would plateau so that we can decide when would be the earliest the evaluation of teachers based on students' heart rate would be meaningful. The next segment reveals the empirical results and the implications of those results.

4. Results & Discussion

From Table II, it can be observed that attributes Up (R=0.5240), Down (R=0.6040) and Peak (R=0.6548) have their correlation significant at the 0.01 level, while attributes Engage (R=0.4270), Time (R=0.3729), Rest (R=0.3895) and Low (R=0.3770) have their correlation significant at the 0.05 level. In simple language, the correlation results point out that heart fluctuations (Up, Down, Peak, and Low) signify classroom engagement and therefore resulted in better scores in the final examination. The average heart rate of the students, which was used to derive the Engage attribute, correlates positively with results of student examination. The amount of time spent in lecture hours in C Programming corresponds positively with the students' grade. However, it is strange that the more time that a student's heart rate is below the resting heart rate baseline during class lecture, the better his or her score will be. It is interesting to note that the Rating attribute of the students on the teacher is neither correlating with their grades nor was the correlation significant. This is a classic example of biasness that is intrinsic in the data collection using questionnaires.

Table 2: R values of the Pearson's Correlation Test of various attributes in relation to SGPA of 32 students in C Programming subject

Attributes	**Pearson's R Values**	**Significance**
Rating	-0.1063	0.56269
Max	-0.0069	0.97007
Min	-0.0357	0.84602
Range	0.0280	0.87909
Engage	**-0.4270***	**0.01480**
Time	**0.3729***	**0.03553**
Stim	-0.3097	0.08450
Rest	**0.3895***	**0.02755**
Slope	0.1481	0.41857
Attend	0.2101	0.24835
Up	**0.5240****	**0.00208**
Down	**0.6040****	**0.00025**
Peak	**0.6548****	**0.00005**
Low	**0.3770***	**0.03341**
*. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed).		

In Table III, the R values are slightly lower. Attribute Peak (R=0.4698) has a correlation significant at 0.01 level while attributes Up (R=0.3873) and Down (R=03925) have their correlation significant at 0.05 level. Since attributes Peak, Up and Down were also shown to be significantly correlating with the SGPA of the students in C Programming subject, it is not surprising that these heart fluctuation attributes contribute to student learning in the classroom setting. Nearing the significant level of 0.05 but not included within the significant level are two attributes, namely Attend (R=0.3379) and Low (R=0.3179). These two attributes may be significant should the sample size increases. It's not a Herculean task to understand the reason these two attributes can be correlated to the SGPA of students. Low is a heart rate fluctuation attribute, which relates with a student's class engagement while Attend is the percentage of class attendance of a student. It can also be seen that the Rating attribute of students were not correlating to their SGPA in any way, neither were attributes like Engage, Time and Rest.

Table 3: R values of the Pearson's Correlation Test of various attributes in relation to SGPA of 32 students in Digital Electronics subject

Attributes	**Pearson's R Values**	**Significance**
Rating	-0.0058	0.97485
Max	0.3258	0.06880
Min	0.1446	0.42983
Range	0.1993	0.27404
Engage	0.1182	0.51935
Time	-0.1014	0.58098
Stim	0.1357	0.45898
Rest	-0.0136	0.94131
Slope	0.1717	0.34744
Attend	0.3379	0.05858
Up	**0.3873***	**0.02853**
Down	**0.3925***	**0.02628**
Peak	**0.4698****	**0.00667**
Low	0.3179	0.07624
*. Correlation is significant at the 0.05 level (2-tailed). **. Correlation is significant at the 0.01 level (2-tailed).		

Observing Table IV, one can testify again that the heart rate fluctuation attributes were obviously correlating with the student SGPA. Rating (R=0.4750), Down (R=0.5435) and Peak (R=0.5431) have a correlation significant at 0.01 level while Up (R=0.4328) and Low (R=0.4245) have a correlation significant at 0.05 level. Interestingly, this is the first time in this study that the attribute Rating correlates positively with the SGPA of students at the 0.01 significance level. However, other attributes such as Engage, Time, and Rest didn't show any relationship, either positive or negative, with the SGPA of the students.

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

Table 4: R values of the Pearson's Correlation Test of various attributes in relation to SGPA of 32 students in Mathematics subject

Attributes	Pearson's R Values	Significance
Rating	**0.4750****	**0.00602**
Max	0.1303	0.47735
Min	0.1137	0.53547
Range	0.0250	0.89216
Engage	-0.2457	0.17531
Time	0.0376	0.83823
Stim	-0.2920	0.10494
Rest	0.1048	0.56808
Slope	0.1295	0.47982
Attend	0.0058	0.97509
Up	**0.4328***	**0.01337**
Down	**0.5435****	**0.00131**
Peak	**0.5431****	**0.00132**
Low	**0.4245***	**0.01546**

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Observing across Table II, III and IV, only attributes Up, Down and Peak were significantly correlating with SGPA of students in all three subjects. Low was significantly correlating with SGPA in two subjects while Rating, Engage, Time and Rest were significantly correlating in one subject. The single best attribute to gauge students' engagement and thus can be used to evaluate teachers' performance is the Peak attribute, followed by Up, Down and Low. Since Mathematics is a revision of what the students learnt in school, it may be the reason that Rating and SGPA correlate. Other attributes like Engage, Time and Rest may correlate coincidently.

From observation of the authors during class lecture, the teacher of Digital Electronics is the strictest while the other two teachers were very friendly. This probably explain why the average Rating and Engage of Digital Electronics is the lowest while the average values of Rest, Up, Down, Peak and Low were the highest. It can be quite stressful to attend Digital Electronics lecture. The Mathematics faculty, on the other hand, is the most experienced faculty and usually tells a lot of stories in class to stimulate the students in learning resulting in a distinct Stim percentage. All these can be seen in Table V. Another observation from the same table is that the average values of Max increases from C to DE to M, the same way of the average values of SGPA. If the trend persists in future research, then Max heart rate can be used to predict the performance of teachers. Even though students attended Mathematics class with the least amount of average time, the percentage of Max, Min, Stim and Engage are the highest and the percentage of Rest is the lowest. It seems students of Mathematics class are averagely tired out at the end of the class with the lowest Slope value. They scored the highest in Mathematics even though their class attendance is the lowest.

From another angle, it can be said that good teachers that produces good academic results among students tend to cause the highest average maximum heart rate of students in class, highest average stimulated percentage, lowest average resting percentage, highest average class engagement percentage, and lowest regression line slope. It seems that the amount of time in class and the attendance percentage aren't critical criteria in analyzing teaching performance.

Table 5: Average values of various attributes collected from 32 students in C Programming, Digital Electronics and Mathematics during 30 hours of class lecture

Attributes	C Programming	Digital Electronics	Mathematics
SGPA	7.2969	7.3750	8.7031
Rating	8.5125	7.9250	8.9375
Max (BPM)	123.8750	124.5625	126.6563
Min (BPM)	61.6875	59.2188	63.8750
Range (BPM)	62.3125	65.4375	62.7188
Engage (%)	63.5938	50.5938	78.9375
Time (second)	3103.9375	3118.8750	2988.6563
Stim (%)	28.9688	26.9063	40.3750
Rest (%)	25.3125	29.1250	17.5313
Slope	-0.0011	-0.0004	-0.0013
Attend (%)	0.8111	0.8176	0.7790
Up	36.9688	41.3750	36.6250
Down	40.7500	44.8125	40.7188
Peak	13.0313	15.9375	13.5938
Low	10.0313	11.9375	10.5000

Fig. 1 is the graph representation of Table V. It is noted that Max is the only attribute having the same pattern as SGPA, where the score increases from C Programming to Digital Electronics to Mathematics.

May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

Therefore, Max may be useful to predict teacher performance as high SGPA score correlates with maximum fun in learning, which is gauged by the maximum heart rate in a class. Other attributes didn't show any similar pattern, whether directly or inversely.

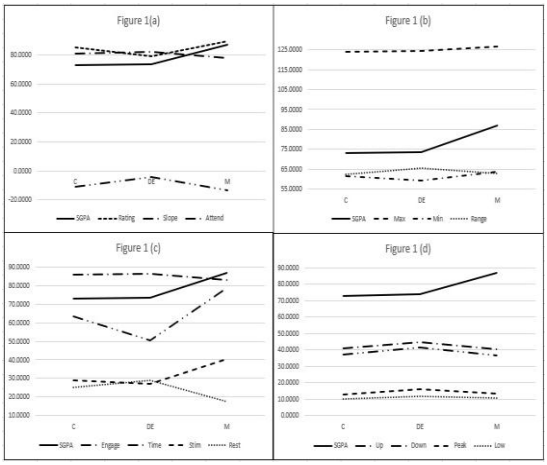

Figure 1: Comparing the average of SGPA of C, DE and M with attributes (a) Rating, Slope and Attend, (b) Max, Min and Range, (c) Engage, Time, Stim and Rest, and (d) Up, Down, Peak and Low.

Since the dataset used in this study is not huge enough to represent the student population, future research should use more health trackers to collect a bigger dataset. Then, more conclusive empirical results might be obtained.

From Fig. 2, one can observe that the cumulative values of the heart rate fluctuation attributes tend to stabilize on Day 13, which is about one month after C Programming class commencement. The same is true in Digital Electronics, as depicted in Fig. 3. However, noting from Fig. 4, it takes a longer time for Mathematics to stabilize, around Day 19. Yet, in all three subjects, none of the average heart rate fluctuation attributes differ more than 3 BPM from Day 1 to Day 30, as the data are arrayed in Table VI. As a result, it can be safely assumed that the average values of Up, Down, Peak and Low collected on Day 1 can be used to evaluate the general performance of teachers. This is especially true for Peak and Low values as they have the least fluctuations in all three ; subjects.

Table 6: Average values of various attributes collected from 32 students in C Programming, Digital Electronics and Mathematics during 30 hours of class lecture

Maximum Difference	**C Programming**	**Digital Electronics**	**Mathematics**
Up	2.53	2.19	1.75
Down	2.88	2.97	1.46
Peak	1.59	1.31	0.91
Low	1.31	2.19	0.97

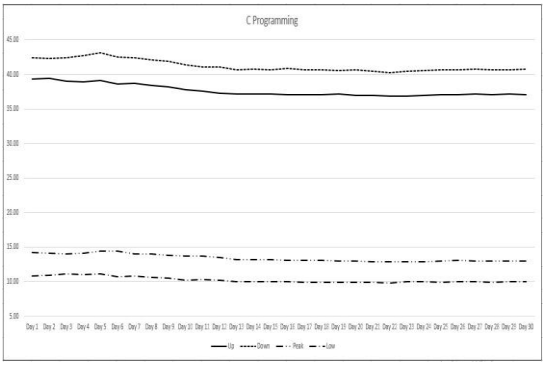

Figure 2: Cumulative values of Up, Down, Peak and Low of 32 students in C Programming subject over 30 lecture hours.

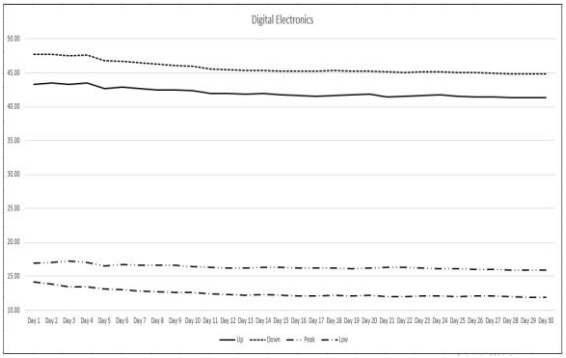

Figure 3: Cumulative values of Up, Down, Peak and Low of 32 students in Digital Electronics subject over 30 lecture hours.

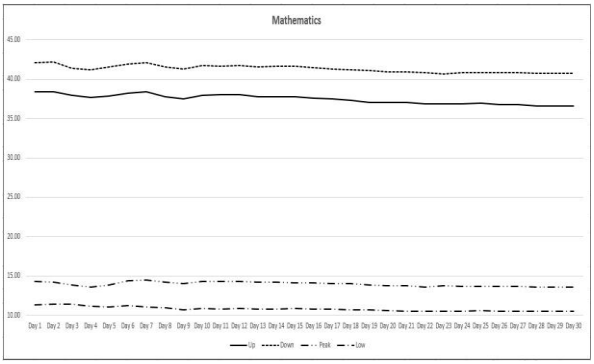

Figure 4. Cumulative values of Up, Down, Peak and Low of 32 students in Mathematics subject over 30 lecture hours.

5. Conclusion

From the empirical results obtained, we can conclude that heart rate fluctuation attributes such as Up, Down, Peak and Low are directly related to the engagement level of the students, which can be used to measure the performance of classroom teaching of each faculty

May-June 2020
ISSN: 0193-4120 Page No. 3753-3759

member. Out of the four attributes, Peak is the most robust and consistent attribute found in all three subjects experimented, namely C Programming, Digital Electronics and Mathematics. Though there are other attributes found to have correlation, but they show correlation only in one subject. Therefore, the significance of such attributes depends on further research.

Of the three subjects, it was noted from the results of Pearson's Correlation analysis that only in the Mathematics was the students' rating of the faculty correlates with the students' SGPA. This pointed to bias that may occur in data collection through questionnaire, as students may fear the consequence of giving an honest feedback. Therefore, there is a pressing need to include physiological data in professional evaluation of teachers, to avoid bias in the analysis.

Generally, the empirical evidence shows that the cumulative values of Up, Down, Peak and Low starting from the first to the last class, didn't fluctuate more than 3BPM. This indicated that these attributes can be used in the beginning to determine which faculty member creates more class engagement.

Besides achieving the objectives of this study, it can be concluded that a good teacher stimulates students in his or her class, shown by high maximum heart rate, high class engagement percentage, and low regression line slope. The students should be tired after class.

This study is limited by the number of students participated in the experiments due to the limited number of devices in heart rate collection. Therefore, the authors look forward to collecting more data and researching other attributes to obtain better analysis parameters to evaluate teaching performance in a non-bias and timely manner.

Acknowledgment

This research is supported by the Vision Group on Science and Technology, Government of Karnataka, under the scheme of Karnataka Fund for Infrastructure Strengthening in Science and Technology Level 1 [GRD No. KSTsPS/VGST-K-FIST L1/2018-2019/GRD NO.788].

References

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man & Cybernetics – Part C: Applications & Reviews, vol. 40, no. 6, pp. 601-618, 2010.

[2] A. Pena-Ayala, "Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works," Expert Systems with Applications, vol. 41, pp. 1432-1462, 2014.

[3] M. Agaoglu, "Predicting Instructor Performance Using Data mining Techniques in Higher Education," IEEE Access, vol. 4, 2016.

[4] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber and M. Nurunnabi, "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation," IEEE Access, vol.7, pp. 108729-108741, 2019.

[5] C.W. Tseng, J.J. Chou and Y.C. Tsai, "Text Mining Analysis of Teaching Evaluation Questionnaires for the Selection of Outstanding Teaching Faculty Members," IEEE Access, vol.6, pp. 72870-72879, 2018.

[6] X. Xu , Y.S Wang and S.J. Yu, "Teaching Performance Evaluation in Smart Campus," IEEE Access, vol.6, pp. 77754-77766, 2018.

[7] D. R. Thompson, J. Di and M. K. Daugherty, "Teaching RFID Information Systems Security," IEEE Transactions On Education, vol.57, no. 1, pp. 42-47, 2014.

[8] B. Fauth, J. Decristan, A.T. Decker, G. Büttner, I. Hardy, E. Klieme and M. Kunter, "The Effects of Teacher Competence on Student Outcomes in Elementary Science Education: The Mediating Role of Teaching Quality," Teaching and Teacher Education, vol. 86, pp. 1-14, 2019.

[9] C. Bradford and M. Braaten, "Teacher Evaluation and the Demoralization of Teachers," Teaching and Teacher Education, vol.75, pp. 49-59, 2018.

[10] S. S. Boswell, "Ratemyprofessors is Hogwash (but I Care): Effects of Ratemyprofessors and University-Administered Teaching Evaluations on Professors," Computers in Human Behavior, vol.56, pp. 155-162, 2016.

[11] Y. Liu, J. Visone, M. B. Mongillo and P. Lisi, "What Matters to Teachers If Evaluation is Meant to Help Them Improve?" Studies in Educational Evaluation, vol. 61, pp. 41-54, 2019.

[12] Haim Shaked, "Why principals often give overly high ratings on teacher evaluations?" Studies in Educational Evaluation, vol.59, pp. 150-157, 2018.

[13] G. Sonnerta, Z. Hazarib and P. M. Sadler, "Evaluating the quality of middle school mathematics teachers, using videos rated by college students," Studies in Educational Evaluation, vol.58, pp. 60-69, 2018.

[14] D. Feistauer and T. Richter, "Validity of Students' Evaluations of Teaching: Biasing Effects of Likability and Prior Subject Interest," Studies in Educational Evaluation, vol.59, pp. 168-178, 2018.