Analysis of Facial Expressions to Estimate the Level of Engagement in Online Lectures

Mrs.S.Gowtham[1], S Dharvesh Mushraf[2], N ArunPrakash[3], S Sridhar[4], K Kavin[5]

[1]Assistant Professor, Computer Science and Engineering Department, Jai Shri Ram Engineering College, Tamil Nadu.

[2,3,4,5] Student, Computer Science and Engineering Department, Jai Shri Ram Engineering, Tamil Nadu.

E-mail: [1]ssgowtham1996@gmail.com, [2]dharveshmushraf@gamil, [3]arunnavalan7@gmail.com, [4]sridhar86101@gmail.com, [5]kavinenegn1806@gmail.com

ABSTRACT

Recognizing and enhancing student engagement is crucial for improving learning outcomes, particularly in the context of online classes where monitoring can be challenging. Traditional methods of attendance tracking, such as calling out names, are impractical and susceptible to manipulation in the virtual environment. This project is to develop a novel design using the AI based FFCNN (Face Fiducial Convolution Neural Network) model to capture face biometric randomly from students' video stream and record their attendance automatically. In the proposed system, the learner's face is monitored by a video camera while attending a video lecture using light Gradient Boosting Machine (LightGBM) to predict the attention and engagement level of the student by facial feature behavior analysis and random QA system.

The present study aimed to develop a method for estimating students' attentional state from facial expressions during online lectures. We estimated the level of attention while students watched a video lecture by measuring reaction time (RT) to detect a target sound that was irrelevant to the lecture. We assumed that RT to such a stimulus would be longer when participants were focusing on the lecture compared with when they were not. We sought to estimate how much learners focus on a lecture using RT measurement. In the experiment, the learner's face was recorded by a video camera while watching a video lecture. Facial features were analyzed to predict RT to a task-irrelevant stimulus, which was assumed to be an index of the level of attention. We applied a machine learning method, light Gradient Boosting Machine (LightGBM), to estimate RTs from facial features extracted as action units (AUs) corresponding to facial muscle movements by an open-source software (OpenFace). The model obtained using LightGBM indicated that RTs to the irrelevant stimuli can be estimated from AUs, suggesting that facial expressions are useful for predicting attentional states while watching lectures. We re-analyzed the data while excluding RT data with sleepy faces of the students to test whether decreased general arousal caused by sleepiness was a significant factor in the RT lengthening observed in the experiment. The results were similar regardless of the inclusion of RTs with sleepy faces, indicating that facial expression can be used to predict learners' level of attention to video lectures.

INTRODUCTION

UNDERSTANDING students' engagement levels while studying is important for improving learning outcomes. To improve the quality of education, it is crucial to estimate learners' level of engagement with their studies. However, it is difficult for teachers to pay attention to all students, particularly in online classes. Automated measurement of engagement levels may be helpful for improving learning conditions. For online learning, webcams can be used to capture learners' facial expressions, which can be used to estimate their mental states . For example, Shioiri et al. conducted image preference estimation from facial expressions and found that this information was useful for estimating subjective judgments of image preference. In education- related studies, Thomas and Jayagopi recorded students' face images in a classroom while they were studying with video material on a screen and estimated the level of engagement from students' facial expressions. The authors succeeded in predicting engagement, suggesting the usefulness of facial expressions for estimating the level of engagement. Heart rate has also been used to estimate mental states during learning. Darnell and Krieg showed that changes in heart rate are related to students' activity during a class [6]. Although previous studies have focused on engagement, which is assessed externally, this research has also been extended to the measurement of internal states, which can be investi- gated by estimating internal states. In these studies, the mental state used as ground truth is based on subjective judgments . However, mental states involve factors other than those that can be evaluated subjectively. Unconscious processes, which cannot be estimated sub- jectively, may play more important roles than conscious

processes. Thus, it is unlikely that subjective judgments are suitable for use as indexes of mental states. For example, heart rate change is reported to be a useful index of students' activity, and is not necessarily related to the subjective estimation of attention and engagement . As such, it is important to develop methods involving objective measures for estimating the level of engagement. A previous study showed that facial features could be useful for estimating reaction time (RT) for mental calculations . This result suggests that RT could be a good index of attention if it varies depending on focusing on the task as typically assumed in attention studies for simple detection, discrimination, or identification of visual stimuli. However, this type of measure is not available for lectures. Therefore, we attempted to use RT for task-irrelevant stimuli.

Although engagement is a term used with different meanings in different contexts , it is often used in relation to attention . Attention to lectures, classes, and tasks is thought to be closely related to engagement. Here we use the term attention to refer to the facilitation of sensory processing by endogenous in- tention or salient exogenous stimulation, and consider it to be a major factor for engagement. It should be noted that engagement has also been used to indicate mental states of a longer duration in some previous studies, such as a whole lecture . We measured levels of attention as an index of engagement during lectures in this study.

We designed an experiment in which participants were asked to detect an auditory target while watching a lecture video. The primary task of the experiment was to understand the lecture, and the secondary task was to detect the target. RT to the auditory target was used as an objective measure of attention level on the lecture videos. Here, we assumed that the time required to detect a target that was irrelevant to the primary task would be longer when the participant focused more on the primary task (i.e., watching video lectures in this experiment). Face images of participants were recorded while watching the videos, and facial expressions were analyzed after the experiment. The purpose of the study was to estimate the RT from facial expressions to develop a method for estimating engagement level from learners' face images.

Some of the results in this study with a smaller num- ber of participants were published in a post-conference book as a preliminary report . Here, we report analyses of facial expressions in more detail with data from a larger number of participants to consider the contributions of specific facial features, the effect of individual variation, and the effect of general arousal level or sleepiness.

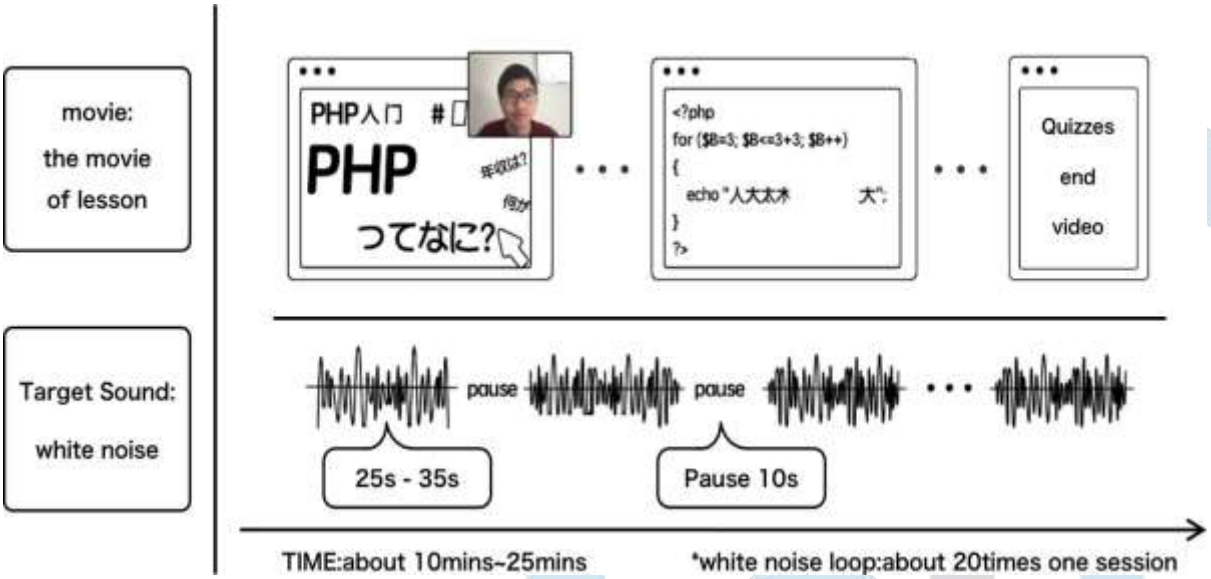
Methodology:

We conducted an experiment to investigate the relation- ship between the attention level and facial expression while watching video lectures. To estimate the level of attention in video lectures,

we measured RT to an auditory target that was irrelevant to the lecture. We assumed that RT to an irrelevant stimulus would be longer when participants were focusing on the lecture compared with when they were not. The effect on brain responses to irrelevant stimuli has been suggested to be able to estimate attention to the primary task. For example, Kramer et al. conducted electroencephalogra- phy (EEG) measurements and reported that the event- related response to a task-irrelevant stimulus changes with the difficulty of a primary task . Similar changes were expected with RT measurements because both ERP and RT have been used to estimate attention in general . In the current study, we attempted to use recorded face images to predict RT.

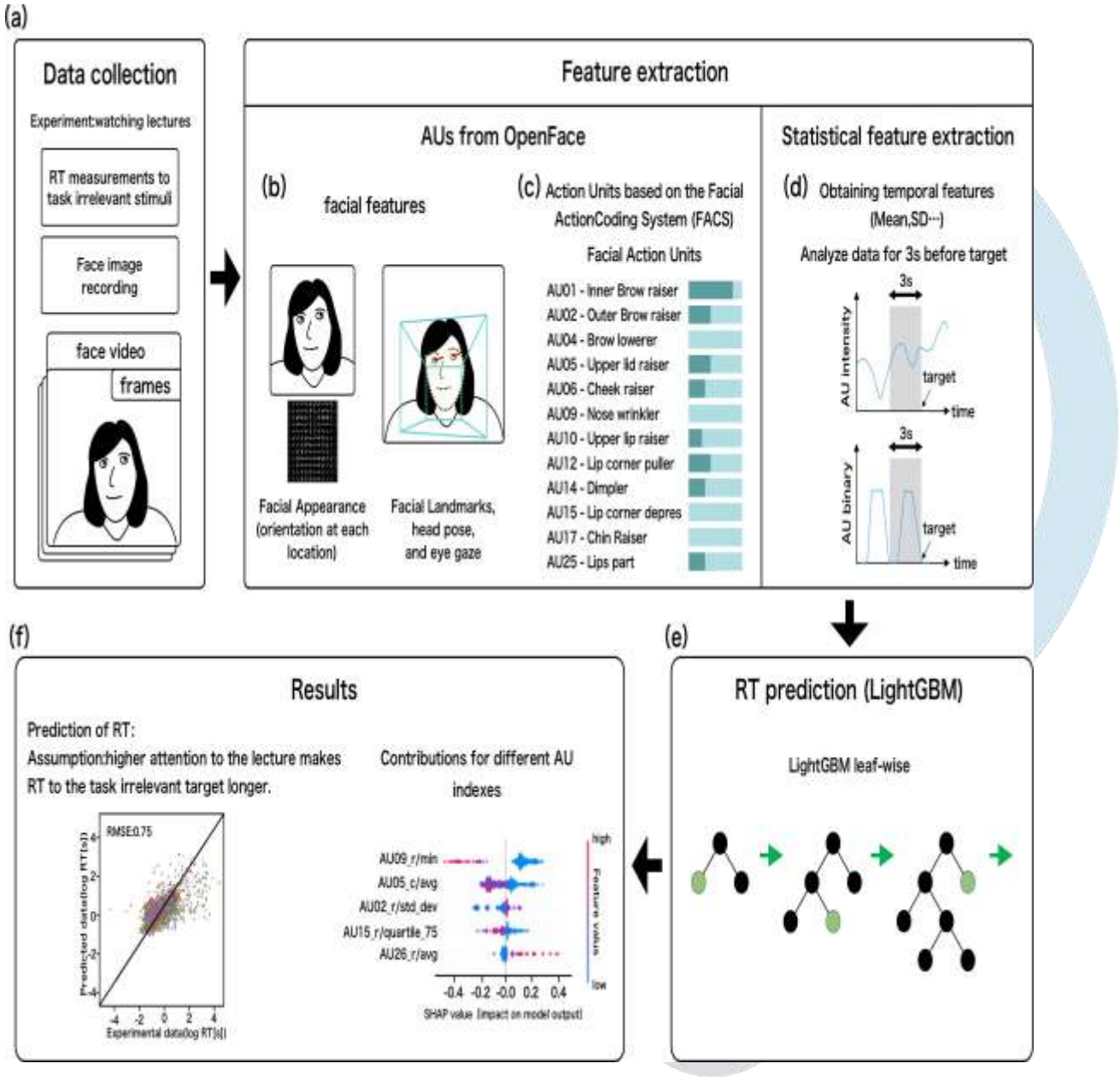
The auditory target we used was the disappearance of continuous white noise instead of the appearance of a sound stimulus, whereas previous experiments to measure attention have typically used a pulse stimulus . The reason for using the disappearance of sound was to avoid the influence of bottom-up attention to a salient stimulus, such as an auditory pulse. Bottom-up attention to a salient stimulus could be strong enough to mask the effect of attention to the lecture. Indeed, the effect of top-down attention cannot be detected when there is only one transient stimulation, while a target is discriminated by top-down attention among many transient stimuli .

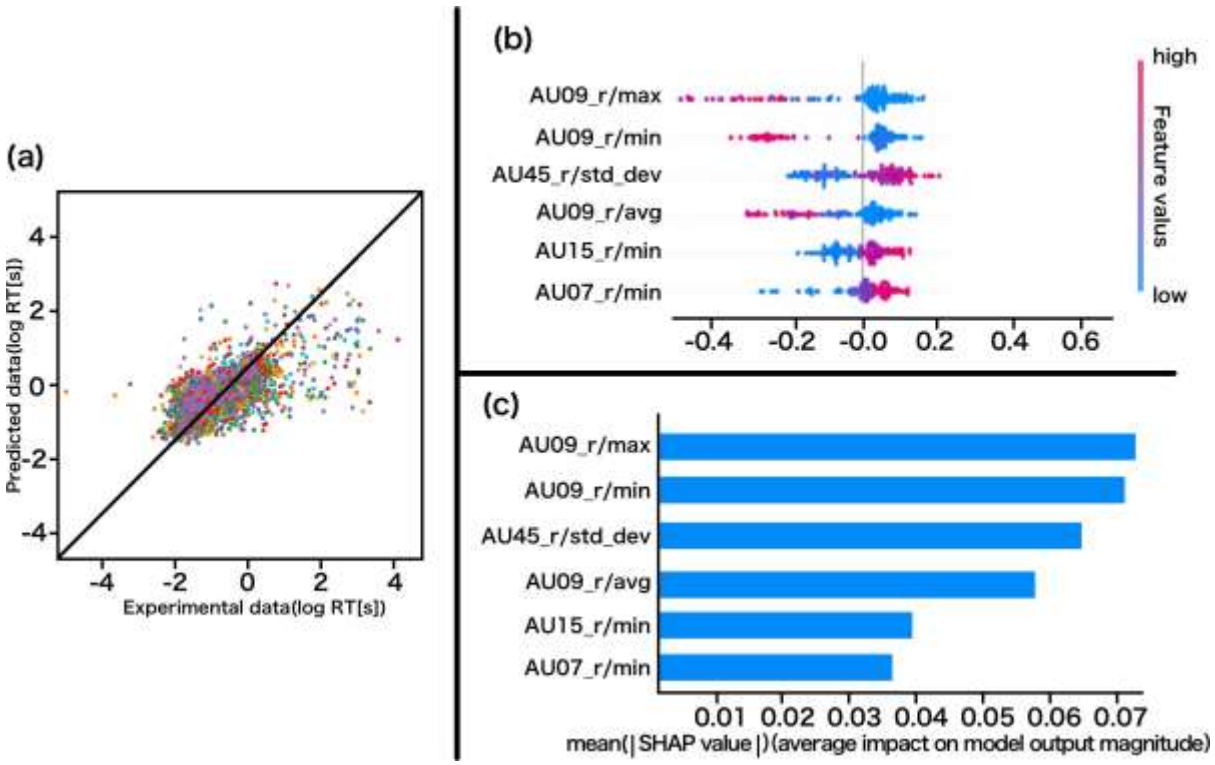
Fifteen participants (average age, 23.1 years) took part in the experiment. Participants had normal or corrected-to- normal vision and normal audition. Par- ticipants were instructed to watch a series of nine video lectures and to answer questions at the end of each video . Participants were also instructed to press a key when they noticed the auditory target (the sudden disappearance of white noise) while watching the video lecture. Participants were instructed that the lecture was the primary task of the experiment while the detection of the target was a secondary task, and they were required to answer questions at the end of the experiment. RT to the target was measured to estimate participants' attention level at the time of the target presentation.



FACIAL FEATURE ANALYSIS

Participants' faces were recorded while watching lecture videos, and their facial features were analyzed after the experiment. We analyzed face images recorded in the 3 seconds before the target presentation (disappearance of white noise), using OpenFace to extract the facial features. To perform facial expression analysis using OpenFace, the first step is to gather facial images or video data. From each video frame, OpenFace detects a face (multiple faces can be detected while there was only one face in our experiment) and locate it in the frame. Then, it makes facial appearance as face orientation and makes facial landmarks such as boundaries of eyes, eyebrows, and mouth. By analyzing the position changes of the facial landmarks and facial appearance, OpenFace evaluates the degree of facial muscle activity as action units (AUs). AUs are assigned to muscle movements related to facial expressions based on the Facial Action Coding System (FACS) . For example, AU1 indi- cates the raising of the inner eyebrows, AU4 indicates the lowering of the eyebrows, and AU5 indicates the raising of the upper lids (Table 1). OpenFace offers several research advantages for facial analysis. Firstly, leveraging deep learning techniques, particularly con- volutional neural networks (CNNs), OpenFace achieves high accuracy in facial recognition and feature extraction


(a)
Data collection
Experiment:watching lectures
RT measurements to task irrelevant stimuli
Face image recording
face video
frames
Feature extraction
AUs from OpenFace
(b)
facial features
Facial Appearance (orientation at each location)
Facial Landmarks, head pose, and eye gaze
(c) Action Units based on the Facial ActionCoding System (FACS)
Facial Action Units
AU01 - Inner Brow raiser
AU02 - Outer Brow raiser
AU04 - Brow lowerer
AU05 - Upper lid raiser
AU06 - Cheek raiser
AU09 - Nose wrinkler
AU10 - Upper lip raiser
AU12 - Lip corner puller
AU14 - Dimpler
AU15 - Lip corner depres
AU17 - Chin Raiser
AU25 - Lips part
Statistical feature extraction
(d) Obtaining temporal features (Mean,SD···)
Analyze data for 3s before target
3s
AU intensity
target
time
3s
AU binary
target
time
(e)
RT prediction (LightGBM)
LightGBM leaf-wise
(f)
Results
Prediction of RT:
Assumption:higher attention to the lecture makes RT to the task irrelevant target longer.
RMSE:0.75
Predicted data(log RT[s])
Experimental data(log RT[s])
Contributions for different AU indexes
AU09_r/min
AU05_c/avg
AU02_r/std_dev
AU15_r/quartile_75
AU26_r/avg
high
Feature value
low
SHAP value (impact on model output)


(a)
Predicted data(log RT[s])
Experimental data(log RT[s])
(b)
AU09_r/max
AU09_r/min
AU45_r/std_dev
AU09_r/avg
AU15_r/min
AU07_r/min
high
Feature value
low
(c)
AU09_r/max
AU09_r/min
AU45_r/std_dev
AU09_r/avg
AU15_r/min
AU07_r/min
mean(| SHAP value |) (average impact on model output magnitude)

Results:

Experimental results indicate that the proposed model achieves high accuracy in engagement classification. The combination of FFCNN and LightGBM outperforms traditional machine learning methods. Challenges such as variations in lighting conditions and occlusions were observed, and possible improvements are discussed.

Target presentations without responses within 10 sec were excluded from the reaction time (RT) analysis. Such target presentations occurred on 5.5% of trials on average across all participants. The average RT over all sessions of all participants was 1.1 sec, with a standard deviation of 2.3 sec. Because average RT varied among participants, we normalized RT as Z-scores after taking the logarithm.

We took the logarithm of RT to minimize the effects of asymmetrical distribution (usually a heavy tail for longer RTs). We also used normalized values of AUs by Z-scoring to avoid the effects of individual variations of facial features. We expected that variations of AUs after normalization were related to changes in mental processes, whereas the absolute AU values include facial differences among different individuals. We then applied LightGBM to model the relationship between RT and facial expressions, and tested the model using a 15-fold cross-validation method. Fig. 3a shows the prediction results of the pooled data model. The horizontal axis shows RT measured in the experiment and the vertical axis shows the prediction from Light- GBM.

Each point represents each target presentation from all sessions of all participants and different colors indicate different training-test combinations (15 differ- ent combinations with different colors). The RMSE of data deviation from the predictions (or the deviation of predictions from the data) was 0.75. The average of the RT data is zero, with a unit standard devi- ation after Z scoring by definition. Thus, the RMSE of model prediction (0.75, which is smaller than 1) indicates that the model can at least partially explain the data variation (25% in this case). The Pearson's correlation coefficient between data and prediction was0.66. A statistical test of no correlation showed that the correlation was statistically significant ($p < 0.001$, $t(2412) = 11$). We used a test to examine whether the Pearson's correlation coefficient is not significantly different from zero and showed the assumption of not different was rejected with a level of 5%.In addition to the statistical significance of correlation coefficient, we also used a statistical test of RMSE to show that our prediction is better than chance. We compared RMSE of the model prediction and that of data, which is one after Z-scoring, using a t-test ($p < 0.001$, $t(14) = 16.62$). The present analysis successfully predicted RT to task-irrelevant targets, which we assumed to vary depending on attention states. This prediction of RT, in turn, predicted the attention state at the time some seconds before the target presentation during learning. We concluded that facial features and movements of the head and eyes contain information about attention.

Conclusion

The present study showed that facial expression can be used to predict learners' level of attention to video lectures, which is an index of students' engagement. It is possible to apply the technology of the use of facial expression to assist in improving the quality of teaching

Facial expression analysis is a viable method for estimating student engagement in online lectures. Our approach demonstrates the potential for real-time engagement monitoring, which can enhance the effectiveness of online education. Future work will focus on improving model robustness and integrating engagement feedback into adaptive learning systems.

REFERENCES

[1] S. Shioiri, Y. Sato, Y. Horaguchi, H. Muraoka, and M. Nihei, "Quali-informatics in the society with yotta scale data," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-4, 2021.

[2] Y. Sato, Y. Horaguchi, L. Vanel, and S. Shioiri, "Prediction of image preferences from spontaneous facial expressions," Interdisciplinary Information Sciences 28 (1), 45-53, 2022.

[3] Y. Horaguchi, Y. Sato, and S. Shioiri, "Estimation of prefer- ences to images by facial expression analysis," IEICE Tech. Rep., vol. 120, no. 306, HIP2020-67, pp. 71-76, 2020 inJapanese.

[4] C. Thomas, D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in Proceedings of 1st ACM SIGCHI International Workshop on Multimodal In- teraction for Education, pp. 33–40, Glasgow, UK, November 2017.

[5] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," Appl In- tell (Dordr), vol. 52, no. 12, pp. 13803-13823, 2022, doi: 10.1007/s10489-022-03200-4

[6] D. K. Darnell, P. A. Krieg, "Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures." PloS one vol. 14,12 e0225709. 2 Dec. 2019, doi:10.1371/journal.pone.