Non-intrusive Identification of Student Attentiveness and Finding Their Correlation with Detectable Facial Emotions

Tasnia Tabassum
Computer Science
Georgia Southern University
Statesboro, GA, USA
tt10589@georgiasouthern.edu

Andrew A. Allen
Computer Science
Georgia Southern University
Statesboro, GA, USA
andrewallen@georgiasouthern.edu

Pradipta De
Computer Science
Georgia Southern University
Statesboro, GA, USA
pde@georgiasouthern.edu

ABSTRACT

Teachers use observational cues in the classroom to identify attentiveness of students and guide the pace of their lecture. However, effectiveness of this technique decreases with increasing class size. This paper presented an approach for automating these observational cues from the students' facial expressions and identifying their attentiveness via a neural network machine learning model. Results of the deep learning Convolutional Neural Network model were then compared with the range of confidence values obtained from a cloud-based emotion recognition service to identify the correlations with human observer. Real time videos of students were collected during classes and sample dataset were created for attentive and inattentive students in order to train the machine learning model. This system can be highly useful for inexperienced teachers for early identification of inattentive students in classroom and thereby taking necessary actions to enhance student learning.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing** → *Education*.

KEYWORDS

Classroom Observational Study, Machine Learning, Convolutional Neural Network, Deep Learning, Amazon Rekognition, Facial Emotion Detection

ACM Reference Format:
Tasnia Tabassum, Andrew A. Allen, and Pradipta De. 2020. Non-intrusive Identification of Student Attentiveness and Finding Their Correlation with Detectable Facial Emotions. In *2020 ACM Southeast Conference (ACMSE 2020), April 2–4, 2020, Tampa, FL, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3374135.3385263

1 INTRODUCTION

Research has shown that engaging students in the learning process tend to increase their focus, encourage them to practice higher-level critical thinking and promote meaningful learning experiences. The most effective classroom evaluations can serve as significant sources of data for instructors for helping them recognize their learning issues and take necessary actions to improve performances [18].

In this paper, we propose a deep learning method for emotional analysis of students in real time classroom environments.In theory, student engagement refers to the degree of attention, curiosity, interest, and passion that students show when they are learning. Different measures can be adopted to determine whether students are fully engaged in the classroom. In [5], the existing methods for learners' engagement detection is divided into three main categories— automatic, semi-automatic, and manual — based on the type of students' involvement in the engagement detection process. The manual methods can be further divided into self-reporting and observational check-list categories. The methods in the automatic category are divided into computer vision analytics, sensor data analysis, and log-file analysis depending on the information that these methods process for engagement detection. The computer vision based methods are further divided into three sub-categories— facial expression based, gestures and postures based, and eye movement based. In the manual category, self-reporting is one of the most popular techniques where a set of questionnaire is posted in which students report their own level of attention, distraction, excitement, or boredom [13]. Observational checklist is another popular method in the manual category for detecting learner engagement that relies on questionnaires completed by external observers. These questionnaires often reflect teachers' personal opinion regarding the learners' engagement levels. Some example questions are "do the learners sit quietly?", "do they do their homework?", "are they on time?", "do they ask questions?" [14]

In this paper, we focused on the automated process of finding student attentiveness by implementing computer vision and machine learning models. The advantage of this method is that it works in the background where the data collection system is passive and therefore does not disrupt the attention span of the students. In other words, there is no observer effect. This method is also gaining popularity because of the ubiquity of the web-cameras in lab computers and classrooms.

Emotions play a critical role in decision-making. However, unlike human, current technologies are incapable of considering human emotions. In this paper, we have come up with a way of finding out whether students are attentive or not in classroom through machine learning model in a way comparable to human observers. There are some challenges while detecting student attentiveness specially in a setup designed to evaluate multiple students at a time. Some students may be talking to other students in the classroom while others may be looking away from the instructor and some others may be looking into their cell phones. Even images of the

ACMSE 2020, April 2–4, 2020, Tampa, FL, USA

ACM ISBN 978-1-4503-7105-6/20/03…$15.00
https://doi.org/10.1145/3374135.3385263

same person with the same facial expression can vary in brightness, background and pose. Variations in shape and ethnicity may also emphasize the inconsistency in the dataset [11]. Therefore, unlike human observers computers might also find it difficult to capture the student emotions and thereby categorize them into attentive or inattentive images at certain times.

Facial expression recognition has been an active research area over the past 10 years, with growing application areas including animation, e-marketing and social robots. The use of Deep Convolutional Neural Networks has attained the best of results in the field of automated facial emotion recognition [15]. In this paper, we propose a methodology for recognizing student behaviour based on facial expression that uses a combination of Convolutional Neural Network and specific image pre-processing steps. However, there are no publicly available dataset with sufficient images for student attentiveness recognition. To tackle this problem, we applied a pre-processing technique to extract the motion sensitive images from video clips of 15 students. We first use a motion detection algorithm to capture students' images and curate a dataset of images for the machine learning model to train on. Domain expert humans were used for labelling images used in training the model. The classifier created thus was able to distinguish between attentive and inattentive students under controlled scenarios. The confidence values returned from the classifier was compared to a set of emotional labels and their respective confidence values to find any underlying correlation.

2 RELATED WORKS

In [22], eye tracker is used to find out students involvement in the classroom by tracking their eye movement. In [8], a non-intrusive model for automated assessment of emotional state along with generation of resulting feedback has been proposed. Geometric features based models trace the shape and size of the facial components such as the eye, mouth or the eyebrows, thereby categorizing the facial expression according to its variation. PCA (Principal Component Analysis) has been used to reduce high dimensional complexity. Local Binary Pattern (LBP) features, a local feature descriptor was used to handle minor changes of facial expressions. The FaceReader [21] uses FACS (Facial Action Coding System) to distinguish six basic emotions with an accuracy of 89%. In [9], authors have achieved considerable accuracy in facial expression recognition by using local features in a person specific dataset. Saccades refer to fast movements of both the eyes in the same direction. Saccades may be recorded using scleral search coils, EOG (Electrooculogram) and high speed videos. Scleral coil and EOG based methods are contact based and hence infeasible for implementations in practical classroom environments. In [20], authors have developed a system to assess human alertness with acoustic stimulation on the basis of the dynamic characteristics of saccadic eye movement. Measurement of PERCLOS is a standard for assessing the alertness level in an individual. PERCLOS is defined as the proportion of time that a student's eyes are closed over a period of time. In [4], authors have investigated the relationship between visual attention and saccadic eye movement. In, [6], Haar-like features have been used to detect the face images. Support Vector Machine (SVM) is used to classify whether the eyes are opened or closed. In the event of PERCLOS exceeding a pre-determined threshold value, a voice alarm is programmed to warn the learner.

Lane et al. [10] conducted observations to note patterns of student behavior that would define engagement or attentiveness. Some of these pre-approved criteria established by experts were used in our research for labeling the dataset.

In [25], the approach uses 2D and 3D data obtained by the Kinect One sensor to build a feature set characterizing both facial and body properties of a student, including gaze point and body posture that portrays attentiveness level in a student. In [24], a Convolutional Neural Network (CNN) based deep learning method is proposed where pre-trained CNN's are utilized for low level feature extraction from each recorded video frame.

In [7], authors present an automated analysis of fine-grained facial movements that occur during computer-based tutoring. They use the Computer Expression Recognition Toolbox (CERT) to track fine-grained facial movements consisting of eyebrow raising (inner and outer), brow lowering, eyelid tightening, and mouth dimpling within a naturalistic video. Within the dataset, upper face movements were linked to be indicators of engagement, frustration, and learning, while mouth dimpling was a positive predictor of learning. CERT finds faces in a video frame, locates facial features for the nearest face, and outputs weights for each tracked facial unit using support vector machine.

In [1], a survey is carried out among instructors involved in e-learning to identify most probable facial features that represent the mood patterns of a student. A neural network approach is then used to train the system using facial feature sets to predict specific facial expressions. Different combinations of inter-related facial expressions for specific time frames were used to estimate facial expressions.The results showed that mood patterns of a student provide a good correlation with his/her involvement during online lectures.

Zhu et al [26] proposed a methodology to identify students' cognitive states using hand motion and heart activity captured with smart watches and applying machine learning model. By applying cross validation, the results of experiments on 30 students in two lectures offered accuracy of 98.99% and 95.78% respectively for predictions of 'interest level' and 'perception of difficulty' for topics covered during the lecture.

In [17], student engagement identification using 3D videos of students while learning, based on facial expression analysis, is proposed. From the facial images, 22 leading landmark points were extracted. The 81 feature vectors constructed from these 22 landmark points, 22 displacement vectors, 16 normalized distance vectors and 21 angle vectors were provided as input to the SVM classifier. The experiment was done on Bosphorus database and also on video database downloaded from YouTube. It gives an average recognition rate of 97.36% for recognizing student engagement using facial expression.

Temporal Convolutional Network (TCN) based framework is used in [19] to understand the intensity of engagement of students attending video tutorials from Massive Open Online Courses (MOOCs). The input to the TCN network is the statistical features computed on 10 second segments of the video from the gaze, head pose and action unit intensities available in OpenFace library. The

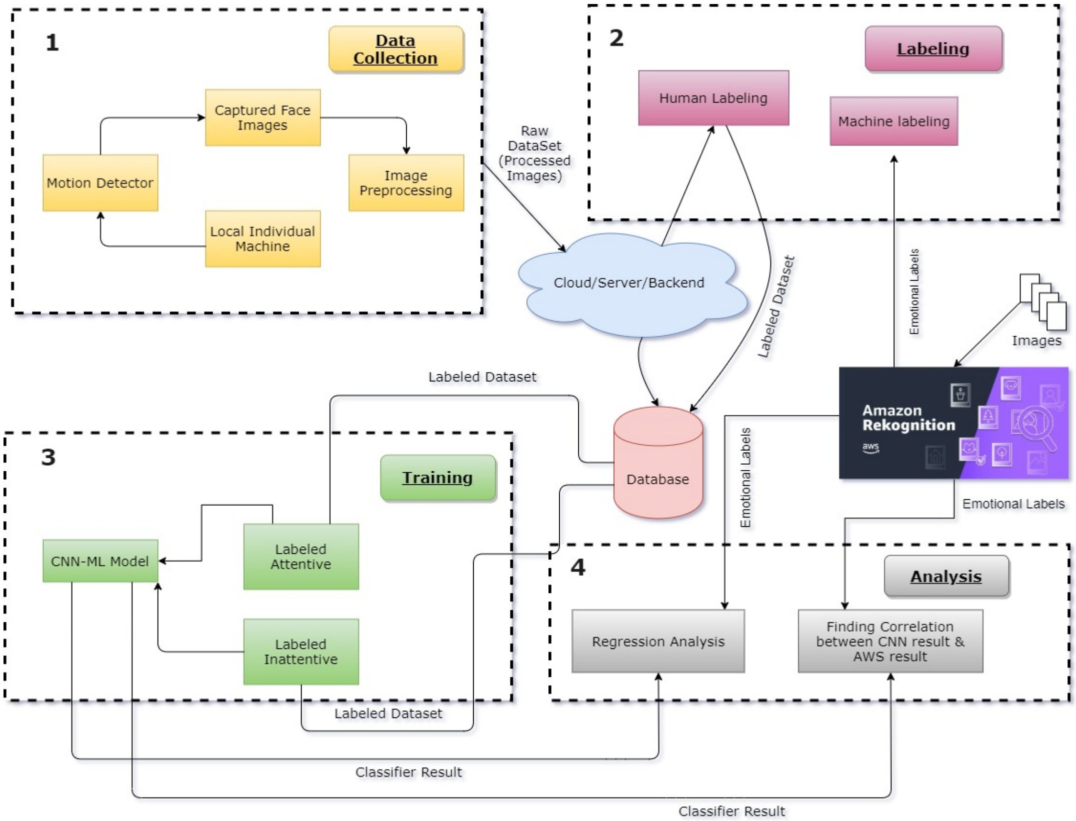


Figure 1: Flowchart of the System Architecture

ability of the TCN architecture to capture long term dependencies provides the ability to outperform other sequential models like LSTMs. In [23], the authors observed that human observers agree when discriminating low versus high degrees of engagement (Cohen’s k ¼ 0:96). Furthermore, they suggested that static expression contains the bulk of the information used by observers. They implemented machine learning to develop automatic engagement detectors and found that for binary classification, automated engagement detectors perform with comparable accuracy to humans. They selected some labelers who were instructed to label engagement based only solely on appearance. The audio was turned off when labeling videos.

In [16] analysis of school classrooms for discriminating between the dimensions “positive climate” and “negative climate” is conducted. Low-level CNN-based facial attribute classifier is trained to detect facial expression and adult vs. child. In addition, a direct audio-to climate regressor integrated with a Bi-LSTM method is implemented to gain an accuracy of 0.40 and 0.51, respectively.

In paper [3], an in-depth comparison of publicly available datasets has been conducted for validating the emotion labels associated with the facial expressions of children. The datasets that were experimented upon are the NIMH Child Emotional Faces Picture Set (NIMH-ChEFS), the Dartmouth Database of Children’s Faces, the Radboud Faces Database, the Child Emotions Picture Set (CEPS), and the Child Affective Facial Expressions Set (CAFE). They compared these systems based on nine attributes such as- age, gender, ethnicity, gaze, geographic location of recruitment, clothing, pose, and number of emotion classes to determine the facial expressions.

[2] compares the results of cloud-based facial expression recognition services offered by Amazon, Google, and Microsoft. Accuracy of the system is measured based on the highest confidence value returned by each of these services. The Karolinska Directed Emotional Faces Database (KDEF) [12] was used to test all three services. The dataset consists of 7 different emotions: afraid, angry, disgusted, happy, neutral, sad, and surprised. Each of the 70 subjects represented each emotion in two different sessions, a total of 140 images per emotion. There were 35 male and 35 female subjects between the ages of 20 and 30. There was no facial hair, visible make-up, eyeglasses or jewelry during the sessions.

However, to the authors knowledge, no study has created a dataset to detect student’s attentiveness and inattentiveness. Motivated by the previous works, in this paper, we designed a non-intrusive attentiveness identification system for students while they were attending lectures and finding a correlation with their corresponding emotional state.

3 METHODOLOGY

The complete image classification pipeline and the corresponding expression analysis of the students for finding a correlation can be divided into four modules (see Figure 1):

- Data Collection Module

- Data Labeling & Pre-processing Module
- Convolutional Neural Network Model Training Module
- Statistical Analysis Module

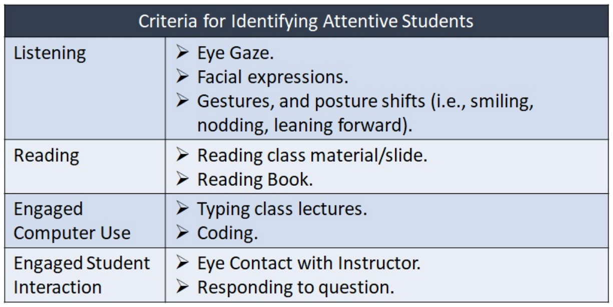
3.1 Data Collection

We captured videos of students while they attend their class lectures. We ran those videos through a motion detector algorithm that captures the images of students from the video whenever a motion is detected. We initialized the size of the area for motion detection threshold. For this step, the equipment setup was 1:1, which means each web camera was dedicated to capture one student's video at a time. The extracted dataset containing raw images were then sent to the next module for pre-processing and labeling.

3.2 Data Pre-processing and Labeling

Figure 2 show the flow for the data pre-processing. The faces of the students are identified using Haar cascade classifiers from the images acquired by the first phase of the pipeline. The images are then cropped and resized into equal dimensions for further processing. The images are then converted to three dimensional image arrays and flattened. CNN learns by continually adding gradient error vectors (multiplied by a learning rate) computed from backpropagation to various weight matrices throughout the network as training examples are passed through. If we did not scale our input training vectors, the ranges of our distributions of feature values would likely be different for each feature, and thus the learning rate would cause corrections in each dimension that would differ from one another proportionally. We might be overcompensating a correction in one weight dimension while under-compensating in another. This is non-ideal as we might find ourselves in an oscillating state or in a slow moving (traveling too slow to get to a better maxima) state. Generally, learning rates are scalars. Thus we normalized the images before using them as input into CNN model. These images are then labeled as either attentive and inattentive by expert individuals for establishing the ground truth for the CNN model. For our research, we considered the attributes or visual cues, presented in Tables 1 and 2, based on Lane et al. [10].

Table 1: Criteria for Identifying Attentive Students

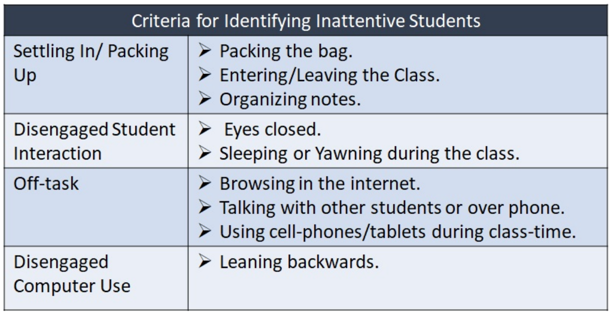


Criteria for Identifying Attentive Students	
Listening	➢ Eye Gaze. ➢ Facial expressions. ➢ Gestures, and posture shifts (i.e., smiling, nodding, leaning forward).
Reading	➢ Reading class material/slide. ➢ Reading Book.
Engaged Computer Use	➢ Typing class lectures. ➢ Coding.
Engaged Student Interaction	➢ Eye Contact with Instructor. ➢ Responding to question.

3.3 Machine Learning Module

3.3.1 Building the CNN model: For input, the CNN classifier takes a normalized set of labeled images, each of which is a three-dimensional matrix. The size of the first two dimensions corresponds to the length and width of the images in pixels and the third dimension corresponds to the channel (1 for gray and 3 for RGB). The CNN comprises of stacks of modules, each of which performs three main operations:

Table 2: Criteria for Identifying Inattentive Students



Criteria for Identifying Inattentive Students	
Settling In/ Packing Up	➢ Packing the bag. ➢ Entering/Leaving the Class. ➢ Organizing notes.
Disengaged Student Interaction	➢ Eyes closed. ➢ Sleeping or Yawning during the class.
Off-task	➢ Browsing in the internet. ➢ Talking with other students or over phone. ➢ Using cell-phones/tablets during class-time.
Disengaged Computer Use	➢ Leaning backwards.

(1) Convolution Layer: The convolution operation is the summation of the element-wise product of two matrices. Convolution (Conv) operation (using an appropriate filter) detects certain features in images, such as horizontal or vertical edges.

(2) Pooling Layer: The pooling layer looks at larger regions (having multiple patches) of the image and captures an aggregate statistic (max, average, etc.) of each region to make the learning invariant to local transformations. The two most popular aggregate functions used in pooling are 'max' and 'average'.
   - Max pooling: If any one of the patches says something strongly about the presence of a certain feature, then the pooling layer counts that feature as 'detected'.
   - Average pooling: If one patch says something very firmly but the other ones disagree, the pooling layer takes the average to find out.

(3) Fully Connected (FC) layer: This layer combines all the local features found in the previous convolutional layers. Each convolution layer computes the output of neurons that are connected to local regions, each computing a dot product between their weights and a small region they are connected to. ReLU (Rectified Linear Unit) layer applies an element-wise activation function, in this case- a Softmax function. The final layer of the neural network has one neuron for each of the classes and they return a value between 0 and 1, which can be inferred as a probability. We can also add dropout in the neural network model to avoid any over-fitting.

Activation Function: We use Softmax as activation function because, in this case, the probability of being one class is not independent of the other class, i.e. the probability of a student being attentive or inattentive should be summed to a total value of 1. It is calculated as follows:

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}$$

Loss Function: Categorical cross-entropy, also called Softmax Loss, is a loss function that is used for single label categorization. This is useful when only one category is applicable for each data point. Therefore, we used categorical cross-entropy here since only one result (attentive/inattentive) can be correct. This will train the CNN

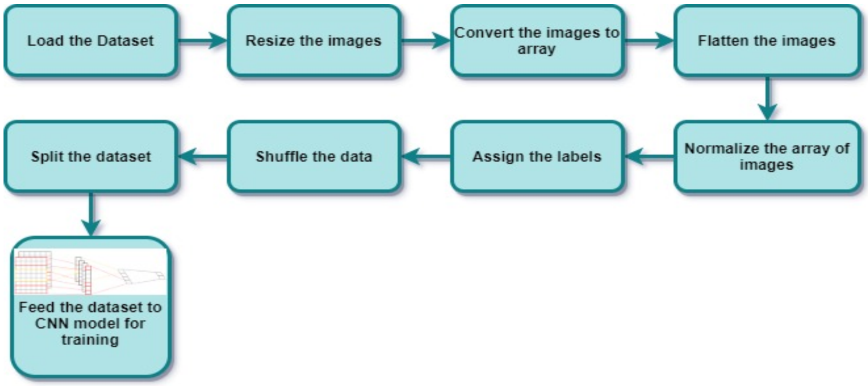


Figure 2: Data Pre-processing for CNN model

to provide a probability output over the C classes for each image. For one-hot encoded labels like ours, only the positive class C_p will keep its term in the loss. It can be calculated as follows:

$$CE = -log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right)$$

3.3.2 Splitting the Dataset: We split the dataset into training and validation dataset with a ratio of 85:15. The training dataset is used for training the CNN model and the validation dataset is used for tuning the hyper-parameters and evaluating the performance of the model.

3.4 Statistical Analysis Module

We ran the images through the Amazon Rekognition system to produce the set of expression analysis for each face image. The facial emotions supported by Amazon's Rekognition API are: happy, sad, angry, confused, disgusted, fear, surprised, and calm. Each image uploaded to the API returns a response containing the analysis results for each face within that given image. Each feature is returned with a confidence value. The confidence value indicates the API's level of confidence in the determined value, it varies between 0 to 100 where all emotion confidence values of an image sums to 100. The result obtained from the classifier trained earlier and the set of emotional values for a specific image is then analysed for finding any correlation.

3.4.1 Correlation Analysis Module: We estimated the Pearson correlation coefficients between attentiveness and different emotions and among the different emotions. This would tell us how differently emotional variables were associated with attentiveness and among themselves. One more reason to calculate the correlation coefficient was to check whether one emotional variable was a linear function of another.

3.4.2 Regression Analysis Module: For the regression analysis, we assumed that expression of attentiveness is a function of the facial emotions mentioned above. In other words,
Expression of attentiveness = *f(facial emotions)*
We have discarded one facial emotions, angry, from the API returned dataset, as anger does not make intuitive sense for our analysis. Expression of this emotion is not prevalent in the undergraduate classroom setting and adding it may cause over-specification of the model. Happy, calm, and surprised are attributed to positive learning experience during learning, while sad, disgusted, confused, and fear are attributed to negative learning experience. However, students' level of attentiveness may be differently influenced than learning experience. We expect disgusted to be negatively correlated with attentiveness, and all the other emotions to be positively correlated to attentiveness. It may appear initially that sadness is negatively correlated with attentiveness; however, because of the similarities between appearing calm and appearing sad, sadness may indicate attentiveness.

We have used an OLS (Ordinary Least Square) regression model to estimate the level of attentiveness. The final model specification was-
Attentiveness = $\beta_0+\beta_1 Disgusted+\beta_2 Calm+\beta_3 Confused+\beta_4 Fear+\beta_5 Happy+\beta_6 Sad+\beta_7 Surprised+\epsilon$
The dependent variable attentiveness was the level of attentiveness provided by the CNN model. The model identifies whether the subjective was attentive and provides the attention level in a 0 to 1 scale, where 1 means completely attentive. However, model does the same for inattentive level, where 1 means completely inattentive, after it identifies the subject as inattentive. To resolve this issue and to run the regression with both dataset simultaneously, we subtracted the inattentiveness level from 1. Now, for both attentive and inattentive, 0 means completely inattentive and 1 means completely attentive. The independent variables are continuous variables as well, ranging from 0 to 1. The regression analysis was performed in Stata/IC 12.0.

4 EXPERIMENTS AND RESULTS

4.1 Data Collection

For conducting the experiment, we collected videos from students during their classes each of which were 30 minutes long. The subjects participating in the experiment were undergraduate students with age ranging between 18-22. Approximately 50% of the subjects were female and it was equally racially distributed. The data were collected in-the-wild, which means it contains data which were recorded using: 1) Unconstrained device (webcamera) placement, 2) Natural environment, 3) Natural behavioral content. The dataset obtained from running the videos through the motion detection algorithm contained around 4000 raw images of students among which some were discarded considering they were noisy and were not very distinctive as to whether they were attentive or inattentive.

4.2 Data Pre-processing and Labeling

The dataset of the raw images consisted of 3500 images. These images were resized to have dimensions of 200*200 (row*column). The entire matrix of image dataset is then converted to arrays and flattened consisting of 3500 rows and 40,000 columns. For classification problems using neural network, it is ideal to perform one hot encoding on the labeled dataset i.e. reshape the output attribute from a vector that contains values for each class label to be a matrix with a boolean for each class value and whether or not a given instance has that class value or not. For example, label "10" means the subject is attentive. The images were then labeled as either attentive or inattentive depending on the criterion mentioned in Figure 1 and Figure 2. Train-test split was then performed on the dataset to acquire 2975 training images 525 validation images. So, the final input to the machine learning module consisted of 1750 attentive images and 1750 inattentive images.

4.3 Machine Learning Model Evaluation

After the training of the model with a batch size of 35 and for 26 epochs, the performance of the classifier reached an optimum level. We used different metrices for evaluating its performance on the validation dataset. The most important of these metrices is the Test Score, Test Accuracy, Classification Report (Table 3), and Confusion Matrix (Table 4). There are few reasons behind it. Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in our dataset- none of these issues are present in our model. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. Therefore, calculating a confusion matrix gives us a better insight on how well our classification model is performing (Table.4). Recall $R = \frac{T_p}{T_p+F_n}$ and specificity $S = \frac{T_n}{T_n+F_p}$ are important measures. F1 score is popular especially in case of binary classification technique.

Table 3: Classification Report for the Classifier Result

	Precision	Recall	f1-score	support
class 0 (ATTENTIVE)	0.93	0.94	0.93	269
class 1 (INATTENTIVE)	0.93	0.93	0.93	256
micro avg	0.93	0.93	0.93	525
macro avg	0.93	0.93	0.93	525
weighted avg	0.93	0.93	0.93	525

$$F1 = \frac{2PR}{P+R}$$

where precision $P = \frac{T_p}{T_p+F_p}$. Tp = true positives, Tn = true negatives, Fp = false positives and F_n = false negatives. Since, our dataset does not suffer from the sensitivity of imbalanced labels, these metrices are sufficient to evaluate our machine learning model. The test score refers to the loss value for the model in test mode which for our model is 0.3993. The test accuracy is 0.9314.

Table 4: Confusion Matrix for the Classifier Result

n= 525	Predicted Attentive	Predicted Inattentive
Actual Attentive	252 (TP)	17 (FN)
Actual Inattentive	19 (FP)	237 (TN)

The graph for visualizing the training accuracy versus validation accuracy is given in Figure 3. The graph is generated for the number of epochs the model was run and the accuracy of the model for each of those epochs. The graph shows that the accuracy increases with each epoch for both the training and the validation dataset which is an indicator of a good model.

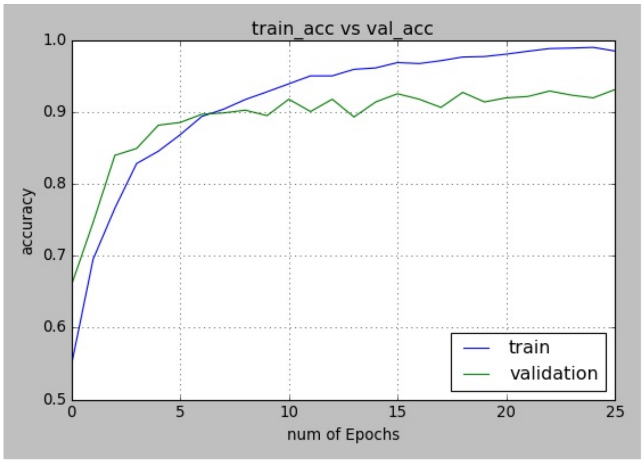


Figure 3: Accuracy Visualization: Training vs. Validation

Note: train_acc and val_acc means training accuracy and validation accuracy, respectively

4.4 Statistical Analysis between Classifier Results and Amazon Rekognition Results

4.4.1 Correlation Analysis: As shown in Table 5, correlation between different emotion variable suggested that attentiveness was positively associated with calm, confused, fear, and surprised, and negatively correlated with disgusted, happy, and sad. Calm was negatively correlated with all the other emotions. Confused was positively correlated with disgusted, and surprised and negatively correlated with other emotions. Fear was positively correlated with disgusted and surprised. Correlation coefficient values among the different emotional variables were low, which suggests low probability of multi-collinearity.

Table 5: Correlation Analysis between different Emotions

	Attentiveness	Calm	Confused	Disgusted	Fear	Happy	Sad	Surprised
Attentiveness	1							
Calm	0.0784	1						
Confused	0.0166	-0.3522	1					
Disgusted	-0.0717	-0.0638	0.0757	1				
Fear	0.1218	-0.2829	-0.0362	0.0728	1			
Happy	-0.2351	-0.2568	-0.0901	0.0672	-0.0394	1		
Sad	-0.0212	-0.5959	-0.1907	-0.038	-0.0143	-0.138	1	
Surprised	0.0395	-0.1436	0.0179	0.1188	0.0526	-0.014	-0.1569	1

4.4.2 Regression Analysis: The regression analysis results (Table 6) indicated that disgusted and happy emotions are negatively affecting attentiveness. While disgusted emotion negatively affecting attentiveness makes intuitive sense, being happy is not so readily understandable. It is possible that the happy emotions are projected from some external source other than from the content of the classroom. However, it is also possible that the model misinterpreted a few occurrences happiness with not being attentive to the class content. By the magnitude of coefficients, disgusted was the strongest predictor of attentiveness. A 1 unit increase in disgusted emotion is followed by a 0.014 unit decrease in attentiveness, all else held constant.

Table 6: Regression Analysis Predicting the Attentiveness of Students

Note: *** and ** indicates significance when $\alpha = 0.01$ and 0.05 respectively

	Coefficients	Standard Error
Disgusted	-0.014***	0.0038
Calm	0.003***	0.0009
Confused	0.003***	0.001
Fear	0.007***	0.0011
Happy	-0.004***	0.001
Sad	0.002**	0.0009
Surprised	0.004***	0.0012
Intercept	0.281***	0.0821

All the variables used in the OLS was found to be statistically significant with an $\alpha = 0.01$, except the emotion sad, which was found significant on $\alpha = 0.05$. Overall model F-statistics was 27.29, which suggests that the model was statistically significant in predicting attentiveness.

5 CONCLUSION AND FUTURE WORKS

In this paper, we presented the traditional neural network based CNN model to classify between attentive and inattentive students in classroom environments and figured out whether there exists any relation between the emotional states of the students and their attentiveness level. One novel contribution of this paper was creating a labeled dataset of attentive and inattentive students attending classes which can be further expanded for new experiments and analysis. The proposed methodology solved the problem of detecting attentive students in classroom with 93% accuracy. With the exception of one facial emotion (happy), we found desired correlation sign for change in attention level. Understanding this correlation will help in identifying the combination of facial expression markers that indicates level of attentiveness and therefore make automating classroom observations easier. This research also helped answer the question whether or not facial expressions are directly linked to student attentiveness. From the results of the analysis, we can conclude that facial emotions are significant predictors of student attentiveness.

One of the challenging tasks of this research was defining student attention in the way to correspond to observations made by the teachers or other human observers or experts and labeling the dataset accordingly since the behavioral metrices defining students as either attentive/inattentive may vary.

In future, we plan to set multiple cameras to record students' interaction with the instructor and observe their change in behavior and sentiments. We also plan to make the dataset bigger and thus increase the accuracy of our prediction result. Such Monitoring system can predict both students' attention over time as well as average attention levels of a class and could be applied as a tool for non-intrusive automated analytics and improvement of the learning process. This system can also be applied in medical sector for surgeons' attention detection in the operation theatre. It can be used to predict drivers' attention level and thereby ensure road safety. Moreover, it can be implemented to study children's interactive learning ability by measuring their attentiveness and correlated emotions.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under grant DUE-1525208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] A. Al-Awni. 2016. Mood Extraction using Facial Features to Improve Learning Curves of Students in E-learning Systems. *International Journal of Advanced Computer Science and Applications* 7, 11 (2016), 444–453.

[2] O. M. Al-Omair and S. Huang. 2018. A Comparative Study on Detection Accuracy of Cloud-Based Emotion Recognition Services. In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*. ACM, 142–148.

[3] D. Bryant and A. Howard. 2019. A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 377–382.

[4] H. Deubel and W. X. Schneider. 1996. Saccade Target Selection and Object Recognition: Evidence for a Common Attentional Mechanism. *Vision research* 36, 12 (1996), 1827–1837.

[5] M. A. A. Dewan, M. Murshed, and F. Lin. 2019. Engagement Detection in Online Learning: A Review. *Smart Learning Environments* 6, 1 (2019), 1.

[6] T. A. Dingus, H. L. Hardee, and W. W. Wierwille. 1987. Development of Models for On-board Detection of Driver Impairment. *Accident Analysis & Prevention* 19, 4 (1987), 271–283.

[7] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. 2013. Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. In *Educational Data Mining 2013*.

[8] S. L. Happy, A. Dasgupta, P. Patnaik, and A. Routray. 2013. Automated Alertness and Emotion Detection for Empathic Feedback during E-learning. In *2013 IEEE Fifth International Conference on Technology for Education (t4e 2013)*. IEEE, 47–50.

[9] S. L. Happy, A. George, and A. Routray. 2012. A Real Time Facial Expression Classification System using Local Binary Patterns. In *2012 4th International conference on intelligent human computer interaction (IHCI)*. IEEE, 1–5.

[10] E. S. Lane and S. E. Harris. 2015. A New Tool for Measuring Student Behavioral Engagement in Large University Classes. *Journal of College Science Teaching* 44, 6 (2015), 83–91.

[11] T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. 2017. Facial Expression Recognition with Convolutional Neural Networks: Coping with few Data and the Training Sample Order. *Pattern Recognition* 61 (2017), 610–628.

[12] D. Lundqvist, A. Flykt, and A. Öhman. 1998. The Karolinska Directed Emotional Faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet* 91 (1998), 630.

[13] H. L. O'Brien and E. G. Toms. 2010. The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69.

[14] J. Parsons and L. Taylor. 2012. *Student Engagement: What do we know and What should we do?* University of Alberta.

[15] G. Pons and D. Masip. 2018. Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. *IEEE Transactions on Affective Computing* 9, 3 (July 2018), 343–350. https://doi.org/10.1109/TAFFC.2017.2753235

[16] A. Ramakrishnan, E. Ottmar, J. LoCasale-Crouch, and J. Whitehill. 2019. Toward Automated Classroom Observation: Predicting Positive and Negative Climate. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[17] R. Ramya, K. Mala, and C. Sindhuja. 2018. Student Engagement Identification based on Facial Expression Analysis using 3D Video/Image of Students. (2018).

[18] K. S. Sahla and T. S. Kumar. 2016. Classroom Teaching Assessment Based on Student Emotions. In *Intelligent Systems Technologies and Applications 2016*, J. M. C. Rodriguez, S. Mitra, S. M. Thampi, and E. El-Alfy (Eds.). Springer International Publishing, Cham, 475–486.

[19] C. Thomas, N. Nair, and D. B. Jayagopi. 2018. Predicting Engagement Intensity in the Wild Using Temporal Convolutional Network. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 604–610.

[20] A. Ueno and Y. Uchikawa. 2004. Relation between Human Alertness, Velocity Wave Profile of Saccade, and Performance of Visual Activities. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 1. IEEE, 933–935.

[21] M. J. Den Uyl and H. Van Kuilenburg. 2005. The FaceReader: Online facial Expression Recognition. In *Proceedings of measuring behavior*, Vol. 30. Citeseer, 589–590.

[22] N. Veliyath, P. De, A. A. Allen, C. B. Hodges, and A. Mitra. 2019. Modeling Students' Attention in the Classroom using Eyetrackers. In *Proceedings of the 2019 ACM Southeast Conference*. ACM, 2–9.

[23] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. R. Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[24] W. Yun, D. Lee, C. Park, J. Kim, and J. Kim. 2018. Automatic Recognition of Children Engagement from Facial Video using Convolutional Neural Networks. *IEEE Transactions on Affective Computing* (2018).

[25] J. Zaletelj and A. Kosir. 2017. Predicting Students' Attention in the Classroom from Kinect Facial and Body Features. *EURASIP journal on image and video processing* 2017, 1 (2017), 80.

[26] Z. Zhu, S. Ober, and R. Jafari. 2017. Modeling and Detecting Student Attention and Interest Level using Wearable Computers. In *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 13–18.