Spring 2019

iFocus: A Framework for Non-intrusive Assessment of Student Attention Level in Classrooms

Narayanan Veliyath

IFOCUS: A FRAMEWORK FOR NON-INTRUSIVE ASSESSMENT OF STUDENT ATTENTION LEVEL IN CLASSROOMS

by

NARAYANAN VELIYATH

(Under the Direction of Pradipta De)

ABSTRACT

The process of learning is not merely determined by what the instructor teaches, but also by how the student receives that information. An attentive student will naturally be more open to obtaining knowledge than a bored or frustrated student. In recent years, tools such as skin temperature measurements and body posture calculations have been developed for the purpose of determining a student's affect, or emotional state ofmind. However, measuring eye-gaze data is particularly noteworthy in that it can collect measurements non-intrusively, while also being relatively simple to set up and use. This paper details how data obtained from such an eye-tracker can be used to predict a student's attention as a measure of affect over the course of a class. From this research, an accuracy of 77% was achieved using the Extreme Gradient Boosting technique of machine learning. The outcome indicates that eye-gaze can be indeed used as a basis for constructing a predictive model.

INDEX WORDS: Affect, Eye-tracking, Machine learning, Attention, Education

IFOCUS: A FRAMEWORK FOR NON-INTRUSIVE ASSESSMENT OF STUDENT ATTENTION LEVEL IN CLASSROOMS

by

NARAYANAN VELIYATH

B.S., Georgia Southern University, 2017

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

STATESBORO, GEORGIA

IFOCUS: A FRAMEWORK FOR NON-INTRUSIVE ASSESSMENT OF STUDENT ATTENTION LEVEL IN CLASSROOMS

by

NARAYANAN VELIYATH

Major Professor: Pradipta De

Committee: Andrew Allen
Charles Hodges

Electronic Version Approved:

May 2019

DEDICATION

This is a work that would not have come to fruition without the support and guidance of so many around me. To Ben, my oldest friend. For always having time to play a game with me, even through your own troubles. To Rana and Scott, my fellow GAs. I could not have made it through graduate school without them supporting me. I genuinely enjoyed the days and late nights we all spent together. To Amulya, for helping me get through one day at a time. Sometimes we forget to stop and smell the roses, and it takes someone else to help us remember. And of course, to my parents, for dragging me kicking and screaming into adulthood. I cannot thank them enough for encouraging me to keep pushing forward even when I gave up on myself. Everything I have done has been thanks to their love and support, and I can only hope that I can make their effort worth it.

ACKNOWLEDGMENTS

I would like to sincerely thank Dr. Pradipta De for his guidance throughout these years. Though I was not always the best pupil, he still patiently helped me along one step at a time. He always acted professionally, and taught me lessons that would be valuable not just for my research but would also serve me when I graduate. It was through his efforts that this research and thesis came to completion. I would also like to thank Dr. Andrew Allen, Dr. Charles Hodges, and Dr. Aniruddha Mitra for their assistance throughout their project. Their knowledge and guidance allowed me to continue to improve in my research.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter will cover the background behind the research conducted for this thesis. It will briefly clarify the history and relevance of the field, and explain why it is of growing importance. It will also introduce how this problem has been researched and studied by others. Finally, the contributions of this research will be explained.

1.1 Motivation

The process of education has remained relatively unchanged for thousands of years. A teacher would stand in front of a group of students and narrate about a subject. The students in turn were expect to sit patiently and absorb the information. It was an exchange designed to be both simple and effective. In more recent years however, this once standard process has begun to undergo change. It has become more apparent that it is just as important to understand how a student learns as much as what they learn.

Bloom’s Taxonomy is a set of models used to deal with three domains of learning: Cognitive, psychomotor, and affective. The cognitive domain deals with knowledge-based training. Psychomotor skills on the other hand, tend to be more physical or action based. The final model set, and the focus of this research, is the affective model. Affect is a term with origins in psychology, referring to a being’s emotions, motivations, and interests. Affective learning is learning that considers these points in addition to the material being studied. While normal education tends to focus on merely *what* is being taught, affective learning additionally chooses to tackle the issue of *how is the student receiving it?*.

This can be considered especially critical because interest is a vital part of learning. An engaged student will not only be more open to gaining knowledge, but is much more likely to retain that information as well. However, this task can be especially challenging because gauging emotions is not a straightforward task. There is no single guide for determining a person’s emotional state of mind. While some types of affect have clear and obvious signals, other affective states may have indicators that differ from person to person. For example, a student resting their head on an arm could be bored, or might simply be used paying attention in that position. Despite having the same physical posture, the affect and meaning behind the action differs greatly.

As there is no standard method for definitively stating a student’s affect, many approaches have been developed to offer potential solutions. Some methods involve a through overview of subjects, taking in measurements such as skin conductive response, body temperature and brainwave readings. These intrusive measures can often be uncomfortable for the user, which in turn can lead to errors in the collected data. Additionally, they often require more physical setup, adding an additional cost to be considered.

As a result, other researchers have turned toward less-intrusive approaches instead. These methods of collection require less active thought from the user, and can still generate valuable results. Body posture analysis, facial analysis, and gaze collection are example of such methods. The chief purpose of these nonintrusive methods is to reduce any bias or errors that may be caused due to discomfort or awareness of the experimental setup. While eye-tracking technology is not a recent innovation, it has become more and more common as a tool in the study of measuring affect.

Eye tracking technology has become increasingly a staple in many virtual reality software, and many computer games have also begun incorporating it. The intent is that by reading and reacting to the users gaze, the user will be able to enjoy a more seamless experience. For research purposes, this intent and reasoning remains largely the same. By allowing the user to act freely, data can be collected from them without hindering them or causing discomfort. The research conducted here focuses on gaze data obtained through an eyetracker.

Gaze data refers to the points in space where a user’s vision is focused. Gaze data is significant in that it has the potential to yield accurate predictions about a user's attention. Perhaps just as importantly, this method is also based around the notion of being almost completely non-intrusive. Furthermore, there is relatively little effort required in setting up such an eyetracker, allowing for relative ease of use. This allows the technology to be readily deployed without significantly impacting any subjects involved in the research. While there are other means of obtaining gaze data, some of them fall into the intrusive collection category.

1.2 Contribution

There were three main outcomes to this research. The first outcome was that data was collected from students in a classroom setting. This was done in order to find students' attention, as well as to gather information that would assist in later predicting that attention. The tool around which this step, and the research as a whole, is focused is the Tobii Eyetracker. Collecting gaze data through this eye tracker is key in that it allows for a non-intrusive approach. Just as importantly, data obtained in this manner can be more easily understood by both the researchers and professors.

The next outcome involved parsing through the collected data. This allowed us to not only view the data that was gathered, but to also transform it into features appropriate for machine learning. This was done by creating software that would not only combine all the data from the various students, but that would also ensure that the data was free from errors. Once the data was cleaned and adjusted, it was then analyzed to find any trends and discrepancies. Finally, the data was transformed into features and collated into a single dataset. This put the information in a manner suitable for machine learning, which was the final step.

The last outcome reached made use of the data collected to create a predictive model of students' attention in a classroom setting. These models represented predictions of a student’s attention, based on the features given. Models were initially developed on a personal level, though later transitioned to an aggregate approach. A peak accuracy of 77% was achieved, which exceeded the results given by a base model. This indicates that it is indeed possible to create accurate predictions of a students attention through information gained from an eyetracker. This demonstrates not only the effectiveness of the models created, but of the experimental setup as a whole.

Figure 1.1: Tobii 4c Eyetracker.

CHAPTER 2

LITERATURE REVIEW

While the research done here focuses on measuring affect through an eyetracker, there are a myriad of other options that have been explored as well. The study of measuring affect has been conducted in areas ranging from standard classrooms to online courses. This chapter will outline some of the other research done in these areas through various means.

2.1 Measuring Affect in Lecture-Based Environments

Lecture-based environments is the term used to refer to a typical classroom setup. In these scenarios, a professor will stand at the front of the classroom and speak to students, often with the aid of presentations. Students generally do not have access to computers or other personal devices in these setups, and are expected to pay attention solely to the instructor and presentation.

In 2005, researchers Slykhuis, Wiebe, and Annetta used a head-mounted eyetracker to capture the gaze data of participants viewing PowerPoint slides (Slykhuis, 2005). Their goal was to determine what parts of the slide drew the most and least attention, so that instructors could develop their content appropriately. Their results revealed that photographs and pictures were generally viewed more often than plain text. However, if the instructor began narration, the focus would then shift to the appropriate text.

A similar experiment was conducted in 2012 by Rosengrant et al (Rosengrant, 2012). This setup used Tobii Glasses to track student attention during a lecture. Their findings showed that students tended to pay more attention to the PowerPoint slides rather than the professor. They also noted that students were more likely to perform *off-task* behaviors if a slide was present.

Bunce et al. chose to use a simpler and more straightforward means of attempting to measure attention (Bunce, 2010). Students in these classrooms would use clickers, small handheld devices with 3 buttons, to indicate when they felt their mind was wandering. Each button represented a different span of time, being less than 1 minute, 1 to 5 minutes, and more than 5 minutes. The results from this experiment indicated that attention patterns occurred cyclically rather than linearly. Previously, one of the theories in education held that attention fell steadily through a class, which these results refuted. It should be noted, however, that the reports generated by the students are inherently prone to bias, and may not be perfectly reliable.

While some researchers choose to use easier modalities, others use more complex methods in the hopes of obtaining better results. For example, EEG signals, which refers to the brain’s electrical activity, can be obtained from a headset. Szafir used this setup to determine a student’s attention during interaction with a robotic instructor (Szafir, 2012). The instructor would narrate a story to the students, and could also change its volume and become more animated if it detected the attention level was too low. Afterwards, the participants were asked questions based on the story told. It was found that participants with a responsive instructor did indeed hold better attention and story recall than participants who did not. This is particularly vital as it demonstrates the effectiveness of real-time feedback in assisting students during a class.

Narrative film watching during a class was also explored to see if mind wandering could be detected (Mills, 2016). For this research, students watching a film would self-report when they felt their mind was not on the movie. Additionally, their eye movements were tracked while watching the film. While mind wandering was able to be reliably captured, it was also found that the local context of the film itself was much more critical in determining gaze patterns compared to reading static text.

Smartwatches are one of the most recent advances in technology. They have also been used as tools to model student’s attention (Zhu, 2017). A single smartwatch device can yield data from not only an accelerometer and gyroscope, but also the heart rate of the wearer. In the experiment where this was used, a strong correlation was found between the features extracted from the gyroscope and the wearer's attention level. In simpler terms, there were strong links between a user's hand movements and how attentive they were feeling.

Body movements have seen considerable use as features for predicting affect. Even just the movement of students' heads was used as the primary feature for an experiment (Raca, 2015). Raca et al. used cameras in a classroom to capture the head movements of multiple students simultaneously. Using an in-class survey to obtain the ground truth, their goal was to model and predict attention using the head motion data. Unfortunately, they found that head motion alone was not sufficient enough data to be used as a predictor of attention. However, they did find correlation between a student's changing states of attention and head movement. They also noted that social cues may also have a significant impact, but were not considered for their research.

Table 2.1: Lecture-Based Environments

Paper	Main Author	Conference	Year	Modality	Type of Affect	Features
Following Student Gaze Patterns in Physical Science Lectures	David Rosengrant	AIP conference proceedings	2012	Tobii glasses	Attention	Areas of interest
Eye-Tracking Students' Attention to PowerPoint Photographs in a Science Education Setting	David A. Slykhuis	Cognition, Technology, and Work	2013	ASL model 501 eyetracker	Attention	Fixations, saccades, areas of interest
How long can students pay attention in class? A study of student attention decline using clickers.	Diane M. Bunce	Journal of Chemical Education	2010	Clickers	Attention	Button presses
Pay attention!: designing adaptive agents that monitor and improve user engagement.	Daniel Szafir	SIGCHI	2012	EEG headset	Attention	EEG data
Modeling and detecting student attention and interest level using wearable computers.	Ziwei Zhu	BSN	2017	Moto 360 smartwatches	Attention, interest	Hand motions, PPG data
Translating head motion into attention-towards processing of students body-language.	Mirko Raca	EDM	2015	Cameras	Attention	Head travel
Automatic gaze-based detection of mind wandering during narrative film comprehension.	Caitlin Mills	EDM	2016	Tobii TX 300 eyetracker	Mind wandering	Fixations, saccades, pupil diameter

2.2 Measuring Affect in Intelligent Tutoring Systems

Intelligent Tutoring Systems, or ITS, refers to computer programs used to assist in the education of students. These systems are notable in that they are designed to give immediate and often personalized feedback to their user. This allows the student to improve their quality of learning, even without access to a human tutor or teacher.

Arroyo et al. used multiple sensors to measure a student's affective states while working with an ITS (Arroyo, 2009). Tools used included cameras, pressure-sensitive computer mice, posture analysis chair, and conductance bracelets. The affective states being measured were interest, excitement, confidence, and frustration Their results found that tutoring systems that made use of the data from the trackers were able to respond better to emotional states that systems that did not. Of particular note was the revelation that affective states were best predicted by the success and progress of the previous problem, rather than the current one.

Sensorless measurement of affect has also been explored. ITS generate log files of users' activity, and these log files can be analyzed to reveal affective states at various times. In 2012, Baker's team used these log files to predict engagement, confusion, frustration, and boredom of students (Baker, 2012). To obtain the ground truth, a form of BROMP was used. BROMP is a process that involves having observers monitor students as they work, and periodically note down their perceived affect. This research found that confusion and frustration were not only uncommon but were also fairly easy to detect. On the other hand, boredom and engagement were trickier to distinguish. This lends credence to the idea that negative affective traits can be more apparent than neutral or positive ones.

Jaques attempted to use eyetracking to predict affects in the MetaTutor ITS (Jaques, 2014). Using features such as fixations, saccades, and areas of interest, his team focused on predicting the emotions of boredom and curiosity. They were not only effective in being able to distinguish between the two emotions successfully, they were also able to correlate certain eye patterns to the affective states as well. Curious students tended to have more fixations and direct saccades, while bored students had more erratic eye patterns.

The MetaTutor ITS was also used for another experiment by Harley et all in 2015

(Harley, 2015). While Jaques focused solely on eyetracking, Harley instead made use of facial recognition, webcameras, and an electrodermal activation sensor. This research focused on a wider range of emotions as well, measuring for over 10 different affective states. Results indicated that the facial analysis had the strongest correlation with the self-reported scores. It was also noted that the EDA sensor had a generally weak agreement with the self-reported scores, meaning it was not as strong a feature as expected.

Table 2.2: Intelligent Tutoring Systems

Paper	Main Author	Conference	Year	Modality	Type of Affect	Features
Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation.	Cristina Conati	Knowledge-Based Systems	2007	Head-mounted eyetracker	Attention	Pupil diameter
Emotion sensors go to school.	Ivon Arroyo	AIED	2009	Pressure mouse, posture analysis chair, camera, conductance bracelet	Confidence, frustration, excitement, interest	Posture, facial expression, skin conductance
Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra.	Ryan d'Baker	EDM	2012	Log files	Boredom, confusion, frustration	Log file data
Predicting affect from gaze data during interaction with an intelligent tutoring system.	Natasha Jaques	IEEE ITS	2014	Tobii T60 eyetracker	Boredom, curiosity	Fixations, saccades
A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system.	Jason Harley	Computers in Human Behavior	2015	EDA skin sensor, camera	Various	EDA activation, facial expressions
Automatic detection of learner's affect from gross body language.	Sidney D'Mello	Applied Artificial Intelligence	2009	Body pressure, camera	Various	Facial expressions, posture patterns
Modeling and understanding students' off-task behavior in intelligent tutoring systems.	Ryan d'Baker	SIGCHI	2007	Log files	Attention	Log file data
Eliciting motivation knowledge from log files towards motivation diagnosis for adaptive systems.	Mihaela Cocea	UMAP	2007	Log files	Motivation, attention	Log file data

2.3 Measuring Affect in Online Courses

Massively Open Online Courses, or MOOCS, have grown in popularity in recent years. These courses tend to primarily take the form of videos and are available only over the Internet. They enable users to not only learn the material at their own pace, but also at their leisure as well. While measuring affect through MOOCs is certainly more of a challenge than in a typical classroom, it is by no means an impossibility.

Sharma et all developed a tool for MOOCs that would notify a student if it felt their attention level was too low (Sharma, 2016). Eye-tracking was used to determine if a student's attention level was appropriate compared to a standard level determined before. If the user's attention dropped below the threshold, the tool would notify the user as to where their attention on the screen should be. Results showed that the number of feedbacks required dropped near the end of the video, showing the effectiveness of the tool. This tool allows for instructors to not only gain an understanding of how users pay attention to the video, but helps students to ensure they are paying attention to the appropriate sections as well.

Sharma also conducted similar research earlier in 2014 (Sharma, 2014). This research was designed to find the effectiveness of attention on learning in MOOCs through the use of eyetracking. SMI RED eyetrackers were used to record participant's gaze while they watched MOOC videos. Additionally, students were divided into categories based on pre- and post-test performance. The features gathered included area of interest misses, fixations, and backtracks. It was found that good learners had less AOI misses, longer fixations, and more backtracks. This indicates that certain features can be used to not only predict attention, but can lead into determine how well a student may do in the class. This in turn could lead to developments in intervention strategies, allowing professors or systems to step in automatically if a student is struggling.

MOOCs have also begun to see use through mobile devices. As such, tools are beginning to be developed for use in this particular area as well. A mobile application was in fact created in 2015 to not only assist students in watching MOOC videos, but to also predict their attention (Pham, 2015). AttentiveLearner, as the application was called, would pause the video if the phone's camera lens was not covered by a finger. Additionally, the camera was used to capture PPG, or heart rate, to use as a feature. The application would also notify the user if it felt their attention had fallen too low for over a few minutes. The models generated were able to predict whether a student was inattentive at a certain period of time.

Engagement is also a critical affective aspect of MOOCs. Similar to attention, it is a measure of how invested in an activity a user is. In 2014, Guo et al conducted a study on determining how video content and setup affects user's engagement (Guo, 2014). Their goal was to determine if how a video was set up was as important to engagement as what was actually taught. To test this, they took data from multiple MOOC courses, including video start and end times, number of times paused, actual video duration, and whether or not post-video questions were attempted. Engagement was measured through the video watched time as well as problem attempts. Their results indicated that shorter videos held more engagement, that videos with heads visible were more engaging than videos without, and that speaking rate also affected engagement. These results show that even in videos, people will focus on how a material is presented just as much as what as being presented.

Table 2.3: MOOCs

Paper	Main Author	Conference	Year	Modality	Type of Affect	Features
How students learn using MOOCs: An eye-tracking insight.	Kshitij Sharma	N/A	2014	SMI RED 250 eyetrackers	Attention	Points of attention, backtracks
A gaze-based learning analytics model: in-video visual feedback to improve learner's attention in MOOCs.	Kshitij Sharma	LAK	2016	Eyetrackers	Attention	Gaze location
AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking.	Phuong Pham	AIED	2015	Mobile phone	Mind wandering	PPG data
How video production affects student engagement: An empirical study of MOOC videos.	Phillip Guo	L@S	2014	MOOC videos	Engagement	Video length, video content
Examining the relations among student motivation, engagement, and retention in a MOOC: A structural equation modeling approach.	Yao Xiong	Global Education Review	2015	MOOC course	Engagement, motivation	Course analysis
Student Emotion, Co-occurrence, and Dropout in a MOOC Context.	John Dillon	EDM	2016	MOOC course	Various	Course analysis
The role of students' motivation and participation in predicting performance in a MOOC.	P.G. de Barba	Journal of Computer Assisted Learning	2016	MOOC course	Motivation	Course analysis
Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course.	Owen Lu	Interactive Learning Environments	2017	MOOC course	Engagement	Course analysis

2.4 Measuring Affect in Electronic Learning Environments

The last area to be introduced is electronic learning environments. These tend to be research and experiments that do not fit into the other categories. The only requirement for being included in this group is that research must be conducted through a computer. As a result, this category encompasses the largest amount of research.

In 2008, Bulger developed a Classroom Behavioral Analysis System software to measure student engagement during class computer sessions (Bulger, 2008). The goal of this research was not to predict or model engagement, but to measure engagement in a computer class. Engagement was defined by determining on-task vs off-task actions. On-task actions included using the appropriate learning materials or related materials, while off-task actions were anything that was deemed not relevant to the given learning objective. This experiment found that engagement was higher as a whole in a class that performed an interactive exercise compared to a class that was lectured to. While this result may seem apparent in hindsight, it does confirm the idea that students who are actively participating will be more attentive and focused than students simply sitting and listening.

Detecting mind wandering during computerized reading is also another avenue commonly pursued. Uzzaman et al in 2011 and Bixler et al in 2016 both performed similar experiments in this area (Bixler, 2016) (Uzzaman, 2011). Both used eyetrackers to obtain eye gaze movements, and use those gaze patterns as features to predict mind wandering. Common features extracted were fixations, saccades, blink duration, and pupil size. While they did find similar results, there were some key differences. Uzzaman utilized self-reports at periodic intervals during the reading. She found strong correlation between certain gaze patterns and the mind wandering reports. On the other hand, Bixler focused more on context-based features involving the text itself. His results showed a link between mind wandering and these features. These experiments show that even minute changes in an otherwise similar setup can lead to different conclusions being reached.

Research into this area has not just been conducted in the civilian sector, but also from a military standpoint as well. In 2011, a framework was proposed for using low cost sensors to assist in training of military trainees (Carroll, 2011). Sensors used in the framework included motion detectors, heart rate monitors, chair pressure sensors, EEG, and eye-trackers. Measured affective states were anger, frustration, boredom, attention, and engagement. The purpose of this framework was to allow for not just better training, but more efficient training as well.

Sensorless approaches have also seen use in this field as well. Keystroke analysis refers to data received from a user's use of keyboard. Features extracted can include pausing behavior, backspace usage, keystroke timing, and relative timing. Bixler et al used this form of analysis to measure boredom, engagement, and neutral affective states for students performing a writing task (Bixler, 2013). The models generated were successful in classifying boredom vs engagement, but had more difficulty in classifying boredom vs neutral affects.

Another sensorless approach was conducted more recently in 2018. Munshi and his team used log files to predict the affective states of students working in an electronic learning environment (Munshi, 2018). One of the key goals in this research was to determine if there was any significant difference in affective states between strong and weaker learners. More specifically, the states measured were boredom, confusion, delight, engagement, frustration, and other. To obtain the ground truth for these affects, BROMP was utilized. They were able to not only find significant correlation between certain affective states and actions taken in prior problems, but were also able to link specific states to *Hi and* Low learners. The *Hi group tended to have more delight values, while the* Low group was found to have more boredom instead.

Even computer games have been used as tools for research in this area. Bosch et al used a physics-based educational computer game as the setting to gather affect of students

(Bosch, 2015). The goal of such a setup was to collect affect in a manner that would feel natural and unlike a controlled lab environment. Webcameras were used to collect facial features, and BROMP was used to obtain the ground truth. The affective states being measured for this research were boredom, confusion, frustration, engagement, and delight. The classifiers had some difficulty in predicting a 5-affect model. Put more simply, the models had low accuracy when trying to distinguish each affective state individually. However, much better results were achieved when affects were predicted one at a time, such as boredom vs others.

In 2015, researcher Park and her team designed an experiment to test the effectiveness of positive emotional feedback through anthropomorphisms on students (Park, 2015). An anthropormorphism is a cartoon-like character with deliberately exaggerated facial features. The facial features would change in response to the student's affective state. Eye gaze data was collected with the Tobii Eyetracker TX300, and features used included fixations/saccades, fixation duration and Areas of Interest. It was found that students who reported more positive affects had better learning outcomes. However, it was also noted that anthropomorphisms did not have an effect on a student's affective state or learning, except where a student had a strong emotional state pre-learning.

Table 2.4: Electronic Learning Environments

Paper	Main Author	Conference	Year	Modality	Type of Affect	Features
Measuring learner engagement in computer-equipped college classrooms.	Monica Bulger		2008	Specialized software	Engagement	Applications used
Automatic gaze-based user-independent detection of mind wandering during computerized reading.	Robert Bixler	UMUAI	2016	Tobbi TX300 eyetracker	Mind wandering	Fixations, saccades, pupil diameter
The eyes know what you are thinking: eye movements as an objective measure of mind wandering.	Sarah Uzzaman	Conscioussness and Cognition	2011	Eyelink 1000 tracker	Mind wandering	Fixations, saccades, pupil diameter
Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits.	Robert Bixler	IUI	2013	Keystroke logger	Engagement, boredom	Keystroke features
Modeling Learners' Cognitive and Affective States to Scaffold SRL in Open-Ended Learning Environments.	Anabil Munshi	UMAP	2018	Log files	Boredom, delight	Log file data
Automatic detection of learning-centered affective states in the wild.	Nigel Bosch	IUI	2015	Webcam	Various	Facial expressions
Modeling trainee affective and cognitive state using low cost sensors.	Meredith Carroll	I/ITSEC	2011	Multiple sensors	Various	Gaze data, PPG, posture analysis
Emotional e-learning through eye tracking.	Marco Porta	EDUCON	2012	Tobii 1750 eyetracker	Various	Fixations, pupil dilation

2.5 Research Comparison

While many of these works do achieve remarkable results, they also tend to have drawbacks as well. For example, while the research done through intelligent tutoring systems is beneficial in that scope, it also has limited use outside of that range. ITS have a very specific domain that they are equipped to teach and respond to. As a result, it can be difficult to take results from that domain and apply it elsewhere. MOOCs are able to receive data from hundreds, if not thousands, of individuals simultaneously. Unfortunately, due to the very nature of the distance-based learning, it is impossible to receive direct affective measurements from the majority of the individuals. This means that most measurement methods must rely on techniques such as self-reported surveys, or analysis of automatically generated log files. This implementations tend to have intrinsic faults that can make the data collected unreliable.

Additionally, many of the intrusive methods employed would not be feasible to implement in a real-world classroom settings. Head-mounted cameras would disturb not only the students wearing it, but most likely also surrounding students and the professor as well. Skin conductance sensors, while effective in their ability to obtain readings from users, also tend to require additional wires and implements. Even modalities such as heart rate monitors, which can appear in a form as simple as a that of a bracelet, have their flaws. Data obtained from devices such as these are not readily understandable. They must be parsed and adjusted to become more readable, which adds an extra layer of work and complexity.

Ideally, research should be conducted in a manner that allows the students to act as freely and naturally as possible, allowing for the best data to be received. In turn, the data received should be effective in what it conveys without being overly complex. The lecture-based environments tended to be have experiments most similar to those that one would expect in a natural educational setting. Additionally, eyetracking technology allows for the collection of data from a student without being intrusive. The combination of these two factors allows for data collection that is both non-intrusive and straightforward to understand.

CHAPTER 3

METHODOLOGY

This chapter will cover the basic experimental design as well as the details of the experiments conducted for this research. The experimental design consists of three main steps, each of which will be briefly introduced. The first stage, which is comprised of the physical setup as well as the software setup of the experiments, will also be expanded on. The physical setup of the experiment, which includes information about the participants and the classroom, will be explained. A general description and clarification of the software used during the data collection process will also be provided. Lastly, the actions undertaken by the volunteers during the experiment will be explained.

3.1 Experimental Design

There are three main stages to this research. The collection phase, the pre-processing or analysis phase, and the model construction phase. Each phase occurs sequentially, as the succeeding step requires the results of the former one. The data gathered from the collection phase passes to the analysis phase. From the analysis phase, the extracted features are sent to the machine learning stage. The results of the machine learning phase are the final outcome of the experiment.

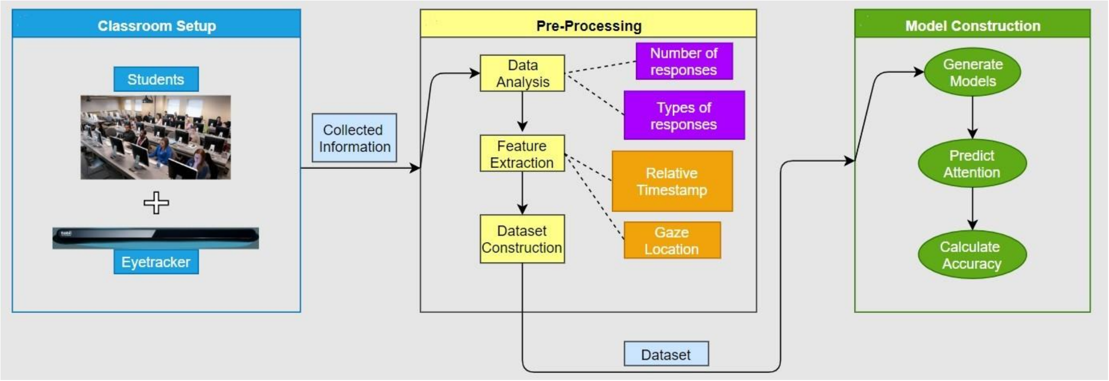

Figure 3.1: Experiment design

The collection phase is the stage where the data was collected from the student volunteers. This is the first stage of the experiment. There are two main components for this phase. The first is the physical setup. This is the aspect of the experiment that deals with the volunteers, hardware setup, and classroom layout. The second is the software portion, which outlines what data was collected, how it was collected, and how the volunteers gave the information. After sufficient data has been collected, it is then sent to the pre-processing stage for analysis and feature extraction.

From the collection phase the data moves to the pre-processing phase. First, the data is analyzed. This yields information such as the number of classes a student attended, or the total number of self-reports generated. This enables us to gain insight into the general state of the class as well as individual students. For example, if multiple students have a low number of self-reports on a given day, that indicates that there could be something happening that is affecting a wide range of students. This additionally has the benefit of allowing outliers to be more easily noticeable. As extreme outliers could be detrimental to the accuracy of the generated models, catching them at this step leads to their removal before the data is turned into features. After all outliers have been identified and removed, the data is processed into features for machine learning. This places the data into predetermined classes for easier comparisons. Finally, the data is stored in a single data set and is ready to be tested on.

The final stage is the model construction phase. This step is where machine learning algorithms are run on the dataset produced by the analysis stage. Different algorithms are used to create the models in order to test a variety of approaches. Models are also trained and tested with different parameters to find the best accuracy. Once the best accuracy has been believed to be found, the results are saved and reported. This concludes one round of the experiment. From here, the experiment resets from the first step and the process begins anew.

3.2 Physical Setup

The participants were volunteers from classes in the field of mechanical engineering. As the experiment was run over two semesters, two separate classes were chosen for testing. The first semester had 12 volunteers in a freshman class of 30. This class met twice a week for 90 minutes each. The second semester has 10 volunteers in a junior class of 25. Out of these 10 volunteers, 6 were male and 4 were female. Unlike the first class, the second class only met once a week, which did limit the data being collected. The professor of the classes had no knowledge of which students were participating in the research, which prevented potential bias. Volunteers signed consent forms and were aware of the purpose and general methodology of the research.

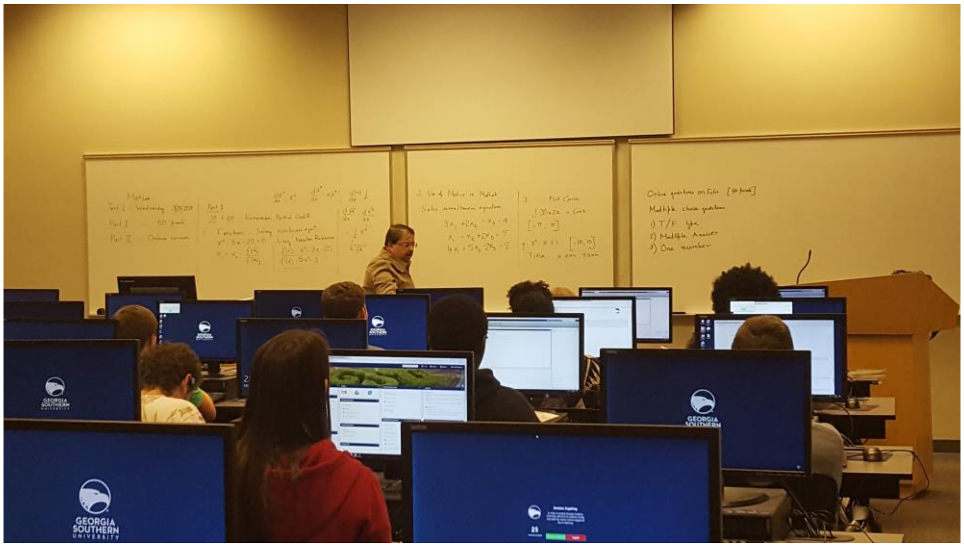

Figure 3.2: Classroom Setup

The classroom was set up in the style of a computer lab. This setup is notably different from a standard classroom in that all students are provided with desktop computers at their desks. In this environment, a professor stands at the front of the classroom, and lectures to the students, often with the aid of PowerPoints or a whiteboard. Meanwhile, students sit at their desks with computers in front of them. They are able to use the computers freely, and also had the option of using physical notebooks and other tools as needed. This setup allows students to either pay attention to a lecture by directly looking at the instructor, or by following allowing with provided material on their computer. In the classroom used, there were 6 computers per row, and 6 rows in total. Eyetrackers were only installed on computers in the first 4 rows.

Eyetrackers were physically secured to the base of each computer monitor and were secured via boltclamps. Each eyetracker was carefully placed so as to not impede nor overly distract any users. No students reported any negative effects from the eyetracker devices during the course of the experiment. All efforts were made to ensure that the eyetrackers did not disrupt the normal classroom process. This is particularly vital, as minimizing intrusion was a key aspect of the research

3.3 Software Setup

For this research, software was implemented to collect, store, and parse through the information collected. The eyegaze coordinates, local timestamp, foreground application name, foreground application coordinates, and ground truth, or self-reported scores, were all gathered and saved to a single text file. This text file was in turn stored on a private server, which was unable to be accessed by any individuals outside of the researchers. Each text file corresponded to a single student on a given class day.

Eyegaze data was obtained from the Tobii eyetrackers. This data was received in the form of 2 coordinates. The coordinates represented the pixel location of the eyes on the computer screen, beginning with (0,0) on the top left and going to the maximum width and height on the bottom right. The gaze location had a margin of error no larger than the size of a quarter. The accuracy of the gaze location was further improved by calibrating the devices to the user’s eyes.

The local timestamp, foreground application name, and foreground application coordinates were obtained directly from the student's computer. This was done through the use of Microsoft APIs. The data collected from these sources occurred in the background and without user intervention or detection. Along with the eyegaze data, this type of information collection is referred to as passive data collection.

To collect the ground truth from the students, a pop-up GUI was generated. This popup can be seen in Figure 3.3. Every five minutes, this pop-up would appear at the top left of the computer screen with a Likert Scale on it. After the **Record** button is pressed, or 1 minute has elapsed, the popup would disappear. This is known as active data collection, as it requires user participation. For the first semester, the scale went from 1-5. For the second semester, the scale was updated to go from 1-10. In both semesters, a 0 or *No Response* option was included.

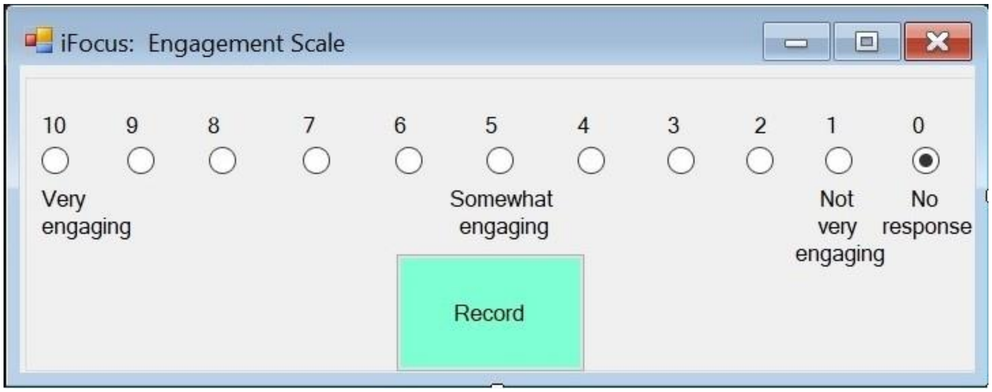

Figure 3.3: Survey Popup

The software additionally performed a check on the user as they logged into the computer. This was done to not only ensure security, but to also make sure that only the necessary data was collected. There were two conditions checked: If the user was in a given list of volunteers; Or if the time/date was appropriate for the correct class periods. If either of these conditions were not met, the software would immediately shut down, and no information would be collected. As the eye-tracker also shut down with the software, it also help prevent the setup from impacting students outside the experiment.

On the first day, volunteers calibrated the software to match their eyes. This was done to ensure that the eyetrackers would match to a specific user, which would increase the accuracy of the gaze collection. Further calibrations were optional. The students were told that every five minutes, a survey would pop up on their screen. They should ideally respond to the survey based on the question asked at the beginning of the experiment. The question asked to the first group of volunteers was “How attentive are you feeling right now?” For the second semester of research, this question was changed to “How engaging was the previous 5 minutes of class?”

CHAPTER 4
DATA PROCESSING

This chapter will go over the data gathered from the collection stage. The first section will cover the analysis of the raw data. This involves results such as the number of responses given, total number of days, and most common scores. The second section will outline the feature extraction process. It will detail the significance of each type of information in relation to predicting attention. It will also note how these data were transformed into features suitable for machine learning.

4.1 Data Analysis

For the first semester, a total of 115 text files were gathered over 11 days. However, 20 of those files had corrupt or unreadable information, leaving them unusable. This resulted in a final total of 95 text files. Of the 12 students that began the experiment, 2 of them left the class before data collection began. Additionally, 1 of the remaining 10 students had a software error, and did not receive any survey popups throughout the experiment. As a result, data for only 9 students was obtained for the first semester. Figure 4.1 below shows the distribution of self-reported scores per student for the first semester.

A total of 684 self-reports were received, not including reports marked as 0. The score of 5 occurred the most often, appearing 328 times. This makes it the most common score by a large margin, being 48% of all scores. In comparison, the next highest score of 3 was only selected 159 times, being 23% of the total makeup. The least common response was 1, which was picked a mere 20 times and had a rate of approximately 3%.

Another key observation was the distribution of responses per volunteer. Students in this class tended to focus on 1 or 2 responses. For instance, student 8 answered 3 a great majority of the time. However, this individual only answered 5 and 0 otherwise. This leads to a large skew in datasets created, which can cause problems during the machine learning

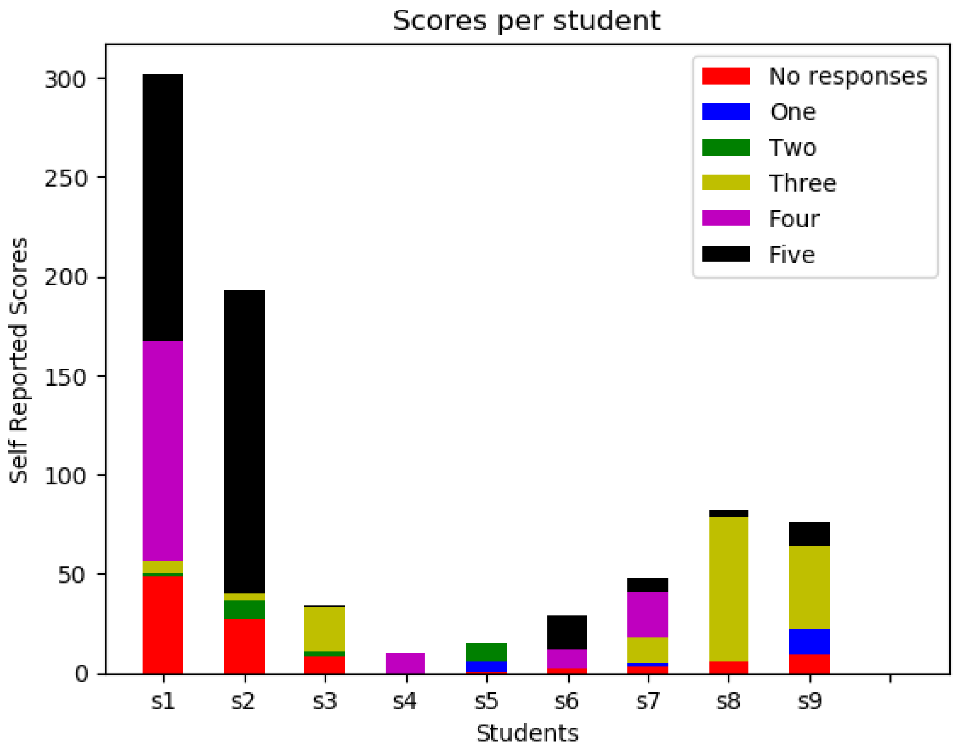

Figure 4.1: First Semester Scores Per Student

phase.

Two key outliers can be quickly identified from this group. Compared to the other volunteers, students 1 and 2 had a considerable number of survey responses. Unfortunately, the exact reasons for this discrepancy remain unknown. The most likely event is that a bug or error caused the popup survey to appear much more often for these individuals compared to the other students. However no factor was found that could cause this incident on only select machines. While attendance of the students was also considered, it was determined that it would not be able to affect the data to such a degree. The results of these 2 students were kept for the final dataset, as otherwise the amount of data collected would be much more limited.

As the second semester met fewer times per week, the amount of data collected is less compared to the first semester. 44 files were recorded over 6 days of class. While all 10 initial volunteers did remain in the class, one of them did not have sufficient survey responses recorded, and was removed from the data set. As a result, the final count led to 41 files were generated by 9 students.

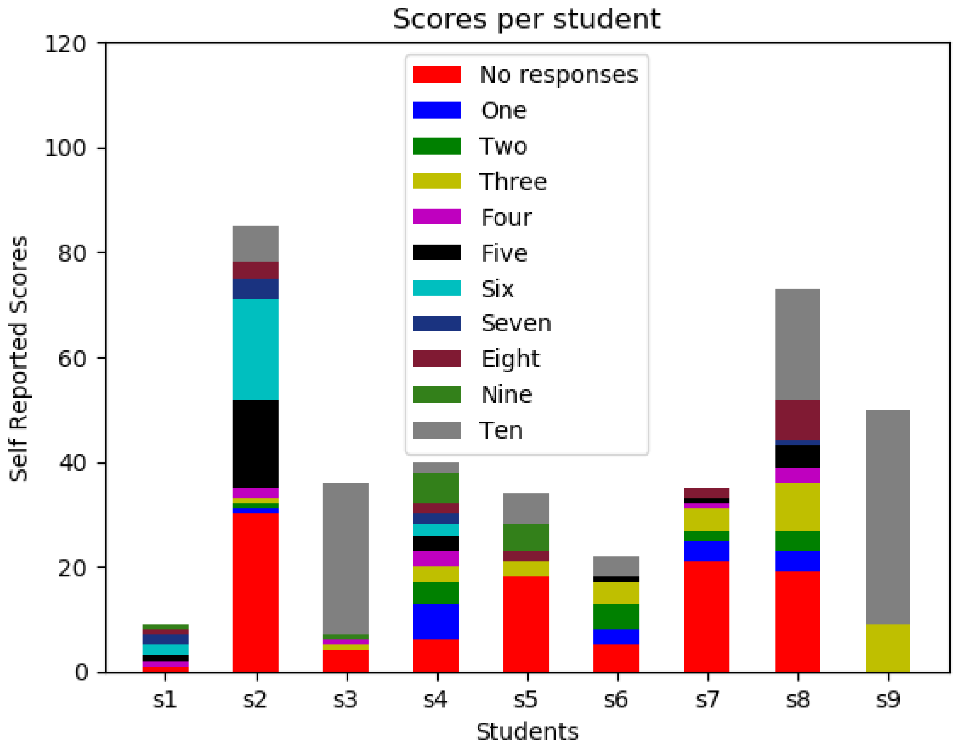

Figure 4.2: Second Semester Scores Per Student

Discounting *No Response* answers, 225 responses were recorded. Out of these responses, 10 was selected most often at 110 times. This is a rate of 49%, which is similar to the response rate of 5 in the previous experiment. For this experiment, the majority of the other responses appeared at a rate similar to each other. Most scores were selected between 11 and 27 times. The lowest responses selected were 3 and 7, occurring only 9 times each.

During this experiment, there was a much more varied distribution of scores per student. Although there was still a large focus on 10 responses, each student also had a respectable number of other responses as well. This may be because the wider scale of 1-10 allows for more natural responses compared to the 1-5 scale used before. The change in question being asked may have also had an impact, as the question was designed to be less personal.

It should be noted that student 9 acted as an interesting outlier. This individual, with the exception of *No Response*, only selected 10 as a response. While this behavior is unusual, it alone cannot be considered as grounds for removal. Without outside knowledge of this student, it might indeed be possible that he/she truly found the class to be completely engaging. As such, this student's data was not remove from the dataset.

4.2 Feature Processing

The eyegaze coordinates, local timestamp, foreground application name, foreground application coordinates, and ground truth, or self-reported scores were collected. Before being fed into machine learning, these values were processed into simpler features. Feature processing is necessary because the open nature of experiment meant that features such as *Applications used* would have an extremely large scope, making comparisons difficult. Features were processed into predefined classes to allow for smoother comparisons.

The eye-gaze coordinates were originally stored as pixel values based on the screen width and height. However, as computer screens can be over 1000 pixels wide, it would be unfeasible to compare those exact values. To rectify this issue, the gaze coordinates were grouped into one of four categories to indicate which quadrant of the screen the student was looking at, This was done in order to determine if there was any correlation between gaze coordinates and attention. The first quadrant was located at the top left, and went counterclockwise, with the fourth and last quadrant situated at the top right.

The foreground application names were also processed. In this case, these values were adjusted to become binary, meaning they could only be 0 or 1. From the data collected, certain websites and applications were noted to be pertinent to school or educational matters. These would be websites such as the class page, PowerPoint slides, and the calculator application. These were labeled as *relevant applications* as they were deemed to be important to the class. All other applications and sites were labeled as *irrelevant applications*. These types of websites included news websites, sports based websites, and video-watching websites. While there was an initial *irrelevant* list, it was quickly discovered that the number of non-relevant applications greatly outpaced the number of relevant ones due to the open nature of the computer network. As such, it would be impossible to list every possible exception, creating the need to adjust the lists into a binary feature. However, there should still remain a strong link between low attention and *irrelevant applications*.

The applications marked as *relevant* were documented through manual input. This was achieved by having a separate program run through a list containing the unique websites and applications visited by each student, and tallying the number of occurrences per application. The resulting lists were then aggregated and then sorted by number of occurrences. The final list was reviewed by hand, and applications deemed to be most relevant to the classroom were inserted into the *relevant* list.

The area of the foreground application was also transformed into a feature. The application coordinates received from each user's computer held the top left and bottom right pixel coordinates. From these values, the area of the application being viewed was able to be calculated. It was believed that this held relevance as there could be a correlation between having multiple windows open and attention. With two windows being open side-by-side, for example, each window would have a small area than a single full size window.

The time stamp was written into the text file as UNIX time. UNIX time is the time elapsed in seconds, or milliseconds for this experiment, since 00:00 1 Jan, 1970. These UNIX values were then converted into local datetime values. The first value when the user first logged in and the software began collecting information was recorded. The time values where the pop-up appeared were also recorded, and the approximate minutes since the log in time was calculated. The difference in minutes was the actual value used for testing.

This difference acts as a relative measure of time elapsed for each user. In a classroom, one would expect that the periods of high and low attentions would be similar for all students, or at least the majority of them.

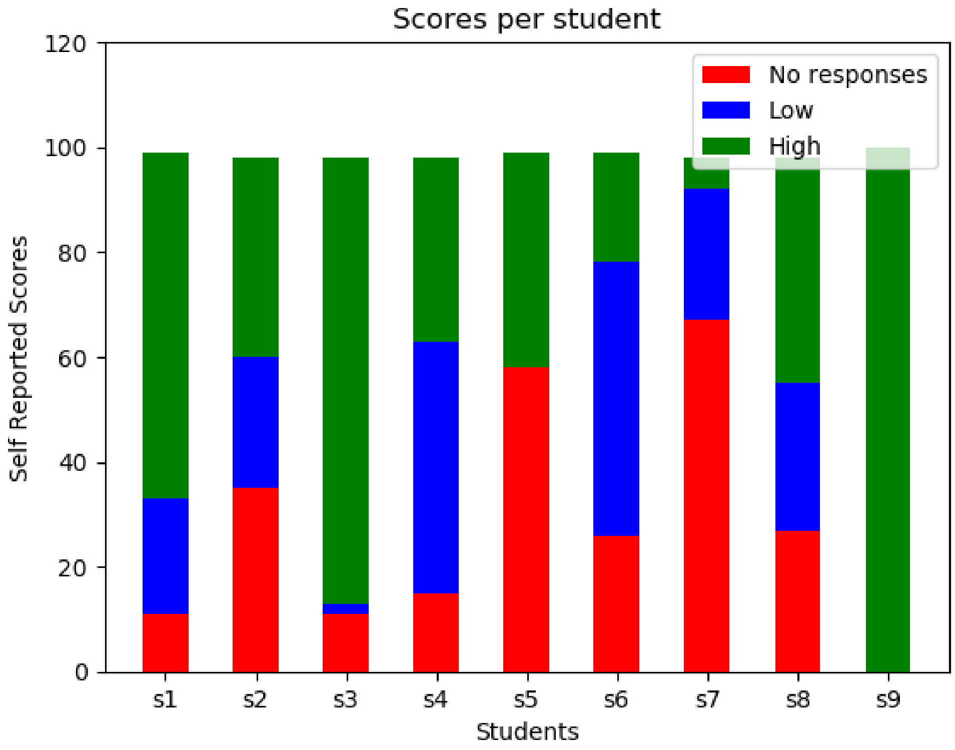

Figure 4.3: Second Semester Scores as Percentages per Student

The last features to be extracted were the self-reported scores themselves. During the first semester of research, the scores were not altered, and were entered into the dataset as is. However, this approach was changed for the second semester. The vital point being predicted was whether or not the student was *attentive* or *not attentive*. The score is a means to find that answer, but the value itself is not a critical component of it. As a result, the students' scores for the second semester of research were transformed to become binary values. Scores of 1-5 were turned into 0, or low attention. High attention was marked as a 1, and was given by reports of 6-10. Figure 4.3 shows how these scores look after being transformed. This figure represents the scores in terms of percent, to show the distribution of scores across **Low**, **High**, and **No Response**. Each student, with the exception of student 9, has atleast some responses in both **Low** and **High** categories. This allows for more accurate and reliable predictions to be made in the machine learning phase.

CHAPTER 5

RESULTS

This chapter will go over the machine learning phase. This is the phase that takes the dataset created from the previous step and attempts to makes predictions on the data. The first section will explain the machine learning methods used. The algorithms and techniques used to generate models will be elaborated on. The metrics used to test the models will also be explained The second section covers the results of the models themselves. The relevant features will also be examined, as well as any other unique tests performed. Finally the best results received will be compared against other contemporary works.

5.1 Models and Metrics Used

Six machine learning models were used over the course of this research: Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), Naive Bayes (NB), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGB). DTs, RFs, NB, and SVM were used for the first research experiment. For the dataset originating from the second semester RF, SVM, Adaboost, and XGB were implemented instead. This change was implemented as it was believed that the new algorithms would be able to provide better and more reliable results than the ones replaced.

Decision trees are one of the simplest classifiers available in machine learning. In this classifier, each feature is tested to lead to a unique outcome. For example, a combination of ‘cloudy,’ ‘wet,’ and ‘warm’ could mean rainy weather, while the choices of ‘cloudy,’ ‘wet,’ and ‘cold’ would instead indicate snow. A decision tree is named as such because the various choices branch out, creating a model that indeed looks like a tree.

Random forests are an extension of the decision tree algorithm. They are created by taking a multitude of decision trees, running them, and then taking the mean of their results. This is done primarily to combat overfitting, which decision trees can be prone to. Overfitting occurs when a model is created that specifically matches the given dataset. While this leads to high accuracy for that dataset, the same model would also fail if given new data.

Naive Bayes was another classifier used for the first dataset. This model performs classification through analysis of the features, then determining the most likely probability from that analysis. A fruit that is red and round would be classified as an apple, while a long yellow fruit would instead be identified as a banana. While this method is simple and effective, it does have a critical weakness. The NB model assumes that all features are independent of each other when performing classification. However, for this research, it is unlikely for this to be true. This is one of the main reasons this algorithm was not used as the research progressed.

Support Vector Machines are not as easy to explain as the previous models, but also tend to have better results. At their core, an SVM simply tries to create a line that separates all classes from each other neatly. It aims to find the best divide, so that the gap between the classes are as large as possible. SVM performs best with binary classifications, although it is capable of handling multiclass problems as well.

Adaboost and XGBoost both belong to the ensemble methods of machine learning. Ensemble methods are more advanced algorithms, used to improve on the results of other models. Adaboost and XGBoost belong to the bagging subset of ensemble methods, which have the effect of reducing innate bias in models. These techniques were chosen because they were simple enough to readily understood, while also having the ability to handle complex data. Adaboost manipulates the weights of weak learning models to allow them to become strong learning models. It does this over multiple trials, and settles on the model that ends up with the best fit. Gradient boosting works on a similar principle, although the details vary significantly. Gradient boosting attempts to minimize loss, which is the difference between the expected and predicted values. It will update predictions to improve performance until the loss function is as minimal as possible. XGB is an extension of gradient boosting which can reach the minimal loss more quickly than the normal version.

The use of neural networks was also considered. A neural network is an advanced framework often containing multiple different algorithms. Inspired by the neural activity in a brain, such artificial neural networks are known for being able to handle extremely complex data. However, neural networks work best when they are fed a large amount of data. As that amount of data was unavailable at the time, neural networks were not chosen as a viable option.

For this problem, there were a few metrics used to test for success. This allowed for the solution to be examined thoroughly. The first and most fundamental metric used was accuracy. Accuracy is the simplest and most straightforward way of judging a model's effectiveness. It is simply a measure of the number of correct responses over the number of total responses. It is typically judged on a scale from 0 - 1.0, where 0 is a complete failure and 1.0 is a total success.

$$Accuracy = \frac{CorrectResponses}{TotalResponses}$$

AUCROC was also used as a testing metric for the second dataset. Standing for Area Under Curve for Receiver Operating Characteristic, AUCROC is used to determine how well a classifier can determine between positives and negatives. When a binary prediction is made, all outcomes can be placed into one of four categories. True positives occur when both the prediction and actual value are positive. On the other hand, true negatives are denoted when the predicted valued and actual value are both negative. If a prediction is negative but the actual value is positive, it is known as a false negative. Lastly, a false positive happens when the actual value is negative but the expected value is positive. AUCROC was chosen because it is known to be better at handling imbalanced datasets.

The metrics of precision, recall, and F-score were also used to aid in the analysis for the second semester's data. Precision is the ability of a model to identify true positives correctly. More exactly, precision checks the number of predictions that are actually correct, or positive, over all predicted positives. Recall, on the other hand, is used to determine how many true positives were found correctly. It can be calculated be taking the number of true positives over all actually positive values. F-score is a combination of precision and recall, and is used to get a quicker overview of both. These metrics were used as they also provide valuable insight into imbalanced datasets. They also offer an additional way to compare the results of various models. These metrics were only used for the second dataset.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

5.2 Results

5.2.1 Machine learning results

For the first dataset, models were built with both personalized and aggregate data. A personalized model meant the predictions were built around a dataset from only one student. An aggregate model on the other hand takes in data from all of the volunteers and makes predictions based off of that. The second semester's dataset only took an aggregate approach, for reasons explained below.

Table 5.1 below shows the results of the personal models created during the first semester. At first glance, the results of the table would seem fantastic. Most of the results are fairly strong and would mean the classifiers worked well. However, a look at each students score distribution reveals the critical flaw in this approach. Most students tend to have an extreme imbalance in their score distribution. This in turn causes problems for measuring accuracy, as models can have trouble with imbalanced datasets. This suspicion was reinforced when examining the aggregate model of the same dataset.

Table 5.1: First Semester Results for Personalized Models

Student	DT	NB	RF	SVM
S1	.49	.57	.48	.49
S2	.84	.97	.86	.93
S3	.78	.87	.83	.87
S4	.93	.93	.93	N/A
S5	.8	.73	.8	.63
S6	1.0	1.0	1.0	N/A
S7	.82	.72	.77	.48
S8	.87	.99	.86	.97
S9	.88	.66	.84	.63

Table 5.2: First Semester Accuracy Results for Aggregate Models

Aggregate	DT	NB	RF	SVM
Accuracy	.46	.48	.46	N/A

As can be seen in Table 5.2, even the aggregate models have very poor results. Even compared to a base model's accuracy of .36, the accuracy of the aggregate models do not even exceed 50%. This is a clear indicator that the dataset has critical and intrinsic issues. As a result, it was deemed that the first semester's data could not be properly reported. It was determined that this data should instead be treated as a trial period, and used instead to ensure no bugs or errors would occur in the next experiment.

For the second semester, only an aggregate approach was taken. This was done not only to combat some of the issues in the first dataset, but also because it was believed that that an aggregate approach would allow for easier feedback to the instructor. Additionally, models created from the second dataset used binary classifiers rather than multiclass ones.

Rather than trying to predict each and every score, the classifiers would instead predict if the attention was high (6-10) or low (1-5). Furthermore, 2 of the previous machine learning algorithms were replaced. Adaboost and XGB were used in place of NB and DT. It was found that NB and DT tended to have weaker results, and so more complex algorithms were selected in the hopes that better accuracy could be achieved. Table 5.3 below shows the best results of these models.

Table 5.3: Second Semester Accuracy Results for Aggregate Models

Aggregate	RF	SVM	Adaboost	XGB
Accuracy	.69	.73	.63	.77

For the second semester, a peak accuracy of .77 was achieved by Extreme Gradient Boosting. On the other hand, the base model only achieved an average accuracy of .52. This is a good indicator that a machine learning model is much more useful than a simple base model. Perhaps just as importantly, it shows a marked difference from the results generated from the first semester. A proper and clean dataset allowed for better results to be produced.

Table 5.4: Second Semester AUCROC Results for Aggregate Models

Aggregate	RF	SVM	Adaboost	XGB
AUCROC	.61	.72	.64	.74

The AUCROC as well as the precision, recall, and F score supports these findings. Tables 5.4 and 5.5 show the best results for each respective value. For AUCROC, XGB once again tended to have the best results. Accuracy measures the effectiveness of a classifier at a certain threshold, while AUCROC instead considers the effectiveness across all thresholds. This means that overall, XGB can be considered to be the strongest classifier of the 4 chosen.

Table 5.5: Precision, Recall and F-Score for each Model

Classifier	Precision	Recall	F-Score
Random Forest	.72	.73	.63
SVM	.78	.975	.87
Ada	.72	.58	.64
XGB	.78	1.0	.88

Once Extreme Gradient Boosting was identified as the best classifier out of the four, parameter tweaking was done to attempt to improve its results. These parameters included the learning objective, learning rate, and base score. The parameter that yielded the most difference was the learning rate. Learning rate refers to the extent of which the weights in a loss gradient function are adjusted. It allows for tuning of the model without needing to adjust any classes or features. However, manipulating the learning rate is not an action without fail. Selecting a learning rate that is too low can result in the gradient descent being slow, taking greater time to find the best accuracy. On the other hand, too high a learning rate can fail to find the best accuracy at all. The graph in Figure 5.1 below shows how the accuracy of the XGB classifier improved as the learning rate was decreased. The accuracy peaked at approximately 77%, with a learning rate of .01. Adjusting this parameter alone allowed for an accuracy increase of approximately 4 percent.

5.2.2 Relevant features

Feature importance was also calculated as part of this research. Feature importance is defined as the impact each feature had on the models' predictions. Figure 5.2 shows a pie chart representing the relative importance of each feature. The feature importance was

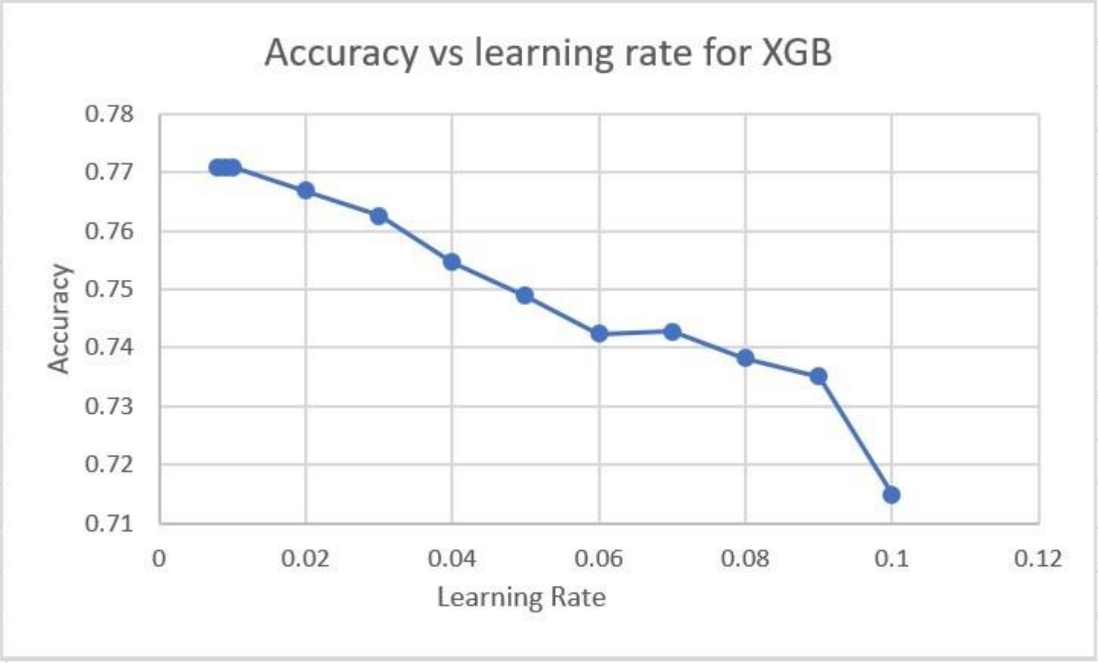

Figure 5.1: Accuracy of XGB Model with Adjustments to Learning Parameter

drawn from the XGB model. As this model had the best results, the features used to make it held greater significance As the figure indicates, timestamps were the most important feature by a decent margin, holding a little under 50% of the importance. This is in line with previous assumptions, as in a classroom setting one would expect many students to have the roughly the same affect at the same times. On the other hand, relevant applications were not as useful as initially thought, being only 9 percent of the total importance. This is most likely due to extreme variance in what each person was looking at, allowing for little correlation between a self-reported score and what they were looking at during that time.

The gaze location was the second best feature, having an importance of .31. This illustrates that gaze features do hold a not-insignificant amount of importance, and do help greatly in predicting attention. While it might not be the best predictor on its own, it can help to supplement other features to produce a better model. The last feature was the application area which held an importance of 14%. While having more of an impact than the application type, it was not nearly as vital as the gaze location or timestamp.

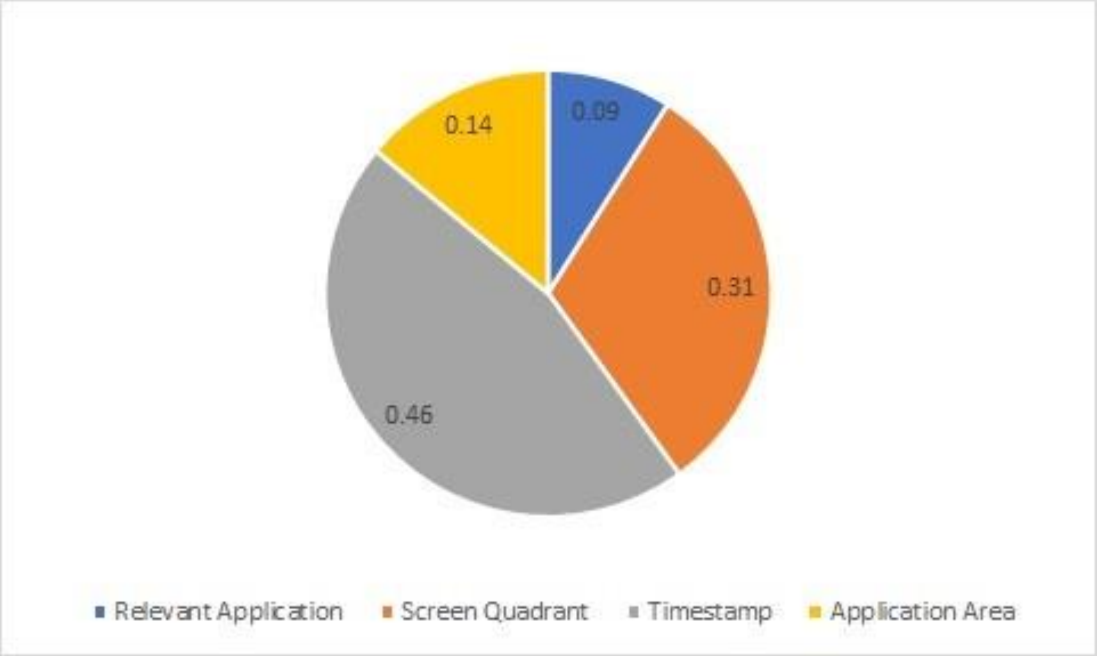

Figure 5.2: Relative Feature Importance

5.2.3 Other tests

In both semesters, a large number of the responses received were marked as 0, or *No Response*. Additional models were created used the second semester's data purely for testing an alternate theory. For these models, the category of 0/**No Response**, was combined into the 6-10/**High Attention** category. This was done under the assumption that students who responded 0 were in fact paying attention, and were responding as such merely to get the survey out of the way. It should be stressed that these models are inherently less reliable, as they are based off a premise that cannot be tested or is fully satisfactory.

Table 5.6: Results of Alternate Hypothesis accepting 0 as High Attention

Aggregate	RF	SVM	Adaboost	XGB
Accuracy	.73	.77	.74	.81

As Table 5.6 shows, the accuracy does increase with this change. XGB remains the strongest classifier for accuracy, achieving an accuracy of over 80%. As previously mentioned, these results exist only for theoretical testing. Due to the initial assumption made when creating these models, the results cannot be claimed to be trustworthy as a predictor. However, this alternate result does open up possibilities for future research. For example, if one could identify the reasoning behind a *No Response* answer, then that could lead to more accurate predictions overall.

CHAPTER 6

LIMITATIONS

While this research did achieve the results it was aiming for, there were some considerable limitations that complicated the experiment. Perhaps the single greatest problem encountered is that there is no way to ensure the ground truth received is 100% accurate. It is entirely possible for students to give false reports out of factors such as boredom or malicious intent. As a result, there is not guaranteed way of knowing from the self-reported scores what a student's real attention level is. As such, all values generated by students must be taken at face value. A possible solution would be to use BROMP, which is a protocol for collecting student engagement. BROMP typically works by having independent observers watch each participant, and note down their apparent affect at certain intervals. While this can offer an unbiased ground truth, it might also affect the classroom proceedings, and would be trickier to implement.

The availability of student volunteers is also a limit. Only certain students will be naturally inclined to volunteer in the first place. For instance, it is unlikely that a student doing poorly in the class would be interested in this experiment. This in turn can cause a skew in the data received. In addition to the ground truths being received, features such as *Relevant Applications* would also be impacted. This issue can be best negated through increased number of trials. As the size of the dataset grows, the influence of each individual student would decrease.

Unfortunately, the Hawthorne Effect cannot ever be mitigated. This effect is used to describe the phenomena where a subject in an experiment will change their behavior, consciously or not, simply due to being aware of the experiment. Without knowledge of a student's behavior outside of the experiment, it is virtually impossible to counter this problem. The nature of a non-intrusive approach means that this problem is as mitigated as much as possible. Additionally, the experiment setup can be adjusted to feel less personal to the users. One such solution already implemented was an alteration to the question being asked for the self-report. By making the question less personal and more about the class, more reliable ground truths can be generated.

Due to being set in an active college classroom, this experiment was also influenced by students outside of the research. It was found that students in other classes would often move or tamper with the eyetrackers. While the boltclamps prevented theft of the eyetrackers, there was no method to prevent them from being moved. These adjustments in turn can cause the eyetracker to return incorrect data or even cease operation. While the eyetrackers were constantly readjusted by the researchers to return them to their proper state as best as possible, this only acts as a temporary cure to the problem. Unfortunately, there does not appear to be a permanent solution to this issue due to the very nature of the setup. It is not possible to monitor and catch all students who tamper with the equipment.

CHAPTER 7

CONCLUSION

7.1 Conclusion

The findings of this research do indeed show that eye gaze data is a useful feature for predicting attention. With a peak accuracy of approximately 77%, the machine learning models exceed a simple baseline model which can only produce an accuracy of 52%. This is fairly in-line with other current works, as the process of machine learning is not a simple one. While gaze data might not necessarily be the best predictor on its own, it is able to work with other features to achieve a reliable level of accuracy. An aggregate model was able to be developed that can easily be returned to and understood by an instructor. While the model may not be able to be created in real time, it stills offers an additional tool for any educator to use. Such a model allows them to look back on a class and determine when and how affect began to change drastically, in turn giving the professor the option to change future classes in response.

7.2 Future Research Directions

There are certainly many avenues that can be taken to further this research. Perhaps the most important, and straightforward, option to take is to add more features. For example, the ability to track and record the professor's behavior would be an excellent feature to include. Such actions would certainly have a great, if not the largest, impact on a student's attention. Additionally, software can be developed to to automatically notify the student if a sufficient amount of time is detected as 'Not attentive'. While this would be a more intrusive approach than the current research, it could also lead to better learning gains for the involved students. Just as students learn from their instructors to gain more knowledge, so too can this research help educators to better understand and help their students.

REFERENCES

[1] Nye, Benjamin, et al. *Analyzing learner affect in a scenario-based intelligent tutoring system.* International Conference on Artificial Intelligence in Education. Springer, Cham, 2017.

[2] Munshi, Anabil, et al. *Modeling Learners' Cognitive and Affective States to Scaffold SRL in Open-Ended Learning Environments.* Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization. ACM, 2018.

[3] Slykhuis, David A., Eric N. Wiebe, and Len A. Annetta. *Eye-tracking students' attention to PowerPoint photographs in a science education setting.* Journal of Science Education and Technology 14.5-6 (2005): 509-520.

[4] Conati, Cristina, and Christina Merten. *Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation.* Knowledge-Based Systems 20.6 (2007): 557-574.

[5] Raca, Mirko, Lukasz Kidzinski, and Pierre Dillenbourg. *Translating head motion into attention-towards processing of students body-language.* Proceedings of the 8th International Conference on Educational Data Mining. No. CONF. 2015.

[6] Bunce, Diane M., Elizabeth A. Flens, and Kelly Y. Neiles. *How long can students pay attention in class? A study of student attention decline using clickers.* Journal of Chemical Education 87.12 (2010): 1438-1443.

[7] d Baker, Ryan SJ, et al. *Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra.* International Educational Data Mining Society (2012).

[8] Lu, Owen HT, et al. *Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course.* Interactive Learning Environments 25.2 (2017): 220-234.

[9] Hutt, Stephen, et al. *The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System.* EDM. 2016.

[10] Park, Babette, et al. *Emotional design and positive emotions in multimedia learning: An eyetracking study on the use of anthropomorphisms.* Computers Education 86 (2015): 30-42.

[11] D'Mello, Sidney, and Art Graesser. *Automatic detection of learner's affect from gross body language.* Applied Artificial Intelligence 23.2 (2009): 123-150.

[12] Harley, Jason M., et al. *A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system.* Computers in Human Behavior 48 (2015): 615-625.

[13] Baker, Ryan SJd. *Modeling and understanding students' off-task behavior in intelligent tutoring systems.* Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2007.

[14] Bixler, Robert, and Sidney D'Mello. *Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits.* Proceedings of the 2013 international conference on Intelligent user interfaces. ACM, 2013.

[15] Cocea, Mihaela, and Stephan Weibelzahl. *Eliciting motivation knowledge from log files towards motivation diagnosis for Adaptive Systems.* International Conference on User Modeling. Springer, Berlin, Heidelberg, 2007.

[16] Whitehill, Jacob, Marian Bartlett, and Javier Movellan. *Automatic facial expression recognition for intelligent tutoring systems.* 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2008.

[17] Mcquiggan, Scott W., Bradford W. Mott, and James C. Lester. *Modeling self-efficacy in intelligent tutoring systems: An inductive approach.* User modeling and user-adapted interaction 18.1-2 (2008): 81-123.

[18] Jaques, Natasha, et al. *Predicting affect from gaze data during interaction with an intelligent tutoring system.* International conference on intelligent tutoring systems. Springer, Cham, 2014.

[19] Wang, Hua, Mark Chignell, and Mitsuru Ishizuka. *Empathic tutoring software agents using real-time eye tracking.* Proceedings of the 2006 symposium on Eye tracking research applications. ACM, 2006.

[20] Bulger, Monica E., et al. *Measuring learner engagement in computer-equipped college classrooms.* Journal of Educational Multimedia and Hypermedia 17.2 (2008): 129-143.

[21] Arroyo, Ivon, et al. *Emotion sensors go to school.* AIED. Vol. 200. 2009.

[22] Szafir, Daniel, and Bilge Mutlu. *Pay attention!: designing adaptive agents that monitor and improve user engagement.* Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2012.

[23] Uzzaman, Sarah, and Steve Joordens. *The eyes know what you are thinking: eye movements as an objective measure of mind wandering.* Consciousness and cognition 20.4 (2011): 1882-1886.

[24] Carroll, Meredith, et al. *Modeling trainee affective and cognitive state using low cost sensors.* Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). 2011.

[25] Porta, Marco, Stefania Ricotti, and Calet Jimenez Perez. *Emotional e-learning through eye tracking."*Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON). IEEE, 2012.

[26] Grafsgaard, Joseph, et al. *Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue.* Educational Data Mining 2014. 2014.

[27] Busjahn, Teresa, et al. *Eye tracking in computing education.* Proceedings of the tenth annual conference on International computing education research. ACM, 2014.

[28] Bixler, Robert, and Sidney DMello. *Automatic gaze-based user-independent detection of mind wandering during computerized reading.* User Modeling and User-Adapted Interaction 26.1 (2016): 33-68.

[29] Afzal, Shazia, and Peter Robinson. *Modelling affect in learning environments-motivation and methods.* 2010 10th IEEE International Conference on Advanced Learning Technologies. IEEE, 2010.

[30] Grafsgaard, Joseph, et al. *Automatically recognizing facial expression: Predicting engagement and frustration.* Educational Data Mining 2013. 2013.

[31] Bosch, Nigel, et al. *Automatic detection of learning-centered affective states in the wild.* Proceedings of the 20th international conference on intelligent user interfaces. ACM, 2015.

[32] Zhu, Ziwei, Sebastian Ober, and Roozbeh Jafari. *Modeling and detecting student at-*

tention and interest level using wearable computers. 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 2017.

[33] Sharma, Kshitij, Patrick Jermann, and Pierre Dillenbourg. *How students learn using MOOCs: An eye-tracking insight.* No. CONF. 2014.

[34] Sharma, Kshitij, et al. *A gaze-based learning analytics model: in-video visual feedback to improve learner's attention in MOOCs.* Proceedings of the Sixth International Conference on Learning Analytics Knowledge. ACM, 2016.

[35] Dillon, John, et al. *Student affect during learning with a MOOC.* Proceedings of the Sixth International Conference on Learning Analytics Knowledge. ACM, 2016.

[36] Pham, Phuong, and Jingtao Wang. *AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking.* International Conference on Artificial Intelligence in Education. Springer, Cham, 2015.

[37] Guo, Philip J., Juho Kim, and Rob Rubin. *How video production affects student engagement: An empirical study of MOOC videos* Proceedings of the first ACM conference on Learning@ scale conference. ACM, 2014.

[38] Xiong, Yao, et al. *Examining the relations among student motivation, engagement, and retention in a MOOC: A structural equation modeling approach.* Global Education Review 2.3 (2015).

[39] Rosengrant, David, et al. *Following student gaze patterns in physical science lectures.* AIP Conference Proceedings. Vol. 1413. No. 1. AIP, 2012.

[40] Chen, ChihMing, JungYing Wang, and ChihMing Yu. *Assessing the attention levels of students by using a novel attention aware system based on brainwave signals.* British Journal of Educational Technology 48.2 (2017): 348-369.

[41] Dillon, John, et al. *Student Emotion, Co-occurrence, and Dropout in a MOOC Context.* EDM. 2016.

[42] De Barba, P. G., Gregor E. Kennedy, and M. D. Ainley. *The role of students' motivation and participation in predicting performance in a MOOC.* Journal of Computer Assisted Learning 32.3 (2016): 218-231.

[43] Mills, Caitlin, et al. *Automatic gaze-based detection of mind wandering during narrative film comprehension.* EDM 16 (2016): 30-37.

[44] McCambridge, Jim, John Witton, and Diana R. Elbourne. *Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects.* Journal of clinical epidemiology 67.3 (2014): 267-277.

[45] Azevedo, Roger, et al. *MetaTutor: A MetaCognitive tool for enhancing self-regulated learning.* 2009 AAAI Fall Symposium Series. 2009.

[46] D'Mello, Sidney, et al. *Gaze tutor: A gaze-reactive intelligent tutoring system.* International Journal of human-computer studies 70.5 (2012): 377-398.

[47] Zaletelj, Janez, and Andrej Koir. *Predicting students attention in the classroom from Kinect facial and body features.* EURASIP Journal on Image and Video Processing 2017.1 (2017): 80.

[48] Smallwood, Jonathan, Daniel J. Fishman, and Jonathan W. Schooler. *Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance.* Psychonomic bulletin review 14.2 (2007): 230-236.

APPENDIX

Purchase and Licensing

The eyetracker model used for this research is a Tobii Eyetracker 4c. The device and information about the device can be located at

https://gaming.tobii.com/product/tobii-eye-tracker-4c/.

For research purposes, a license for the device must also be purchased. This license is necessary to retrieve eye-gaze data from the eye-tracking advice. Information on the license can be found at

https://help.tobii.com/hc/en-us/articles/210251205-Can-I-use-a-Tobii-Eye-Tracker-for-research-purposes-

Tobii Eyetracker Software Installation Process

1. Download the Tobii Core Software. Software can be located at https://gaming.tobii.com/getstarted/
2. The download will appear under the name 'Tobii Eye Tracking Core v2.13.4.7864 x86', with the version number dependent on the most current version. Once the download has completed, click the file to run the installation process for the Tobii Core Software.
3. Complete the installation process for the Tobii Core Software.
4. Plug in the Tobii Eyetracker into a USB port on the computer. The Tobii Core Software should automatically detect the new device and update drivers as necessary.
5. Download the Tobii Pro EyeTracker Manager. This software can be located at https://www.tobiipro.com/product-listing/eye-tracker-manager/

6. The download will appear under the name ‘Tobii.Pro.Eye.Tracker.Manager.Windows-AMD64-1.12.1’. Once the download has completed, click the file to run the installation process for the Tobii Pro EyeTracker Manager.
7. . Complete the installation process for the Tobii Pro EyeTracker Manager. Once the process is complete, click the Manager icon to open up the application.
8. Through the EyeTracker Manager, attach the license to the connected Tobii Eye-tracker.
9. The Eyetracker installation process is now complete.