Classifying Facial Expressions of Students Being Tutored in a Gateway College Math Course

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Kriss Gordon Gabourel

May 2024

This work is dedicated to my mother, who was my very first tutor and who taught me to never stop being curious. I love you and miss you...

Acknowledgments

I give You praise, O Lord, because *"You have granted me eternal blessings and made me glad with the joy of your presence"* (Psalm 21:6)

This absolutely would not have happened were it not for the encouragement and support of so many. I would like to offer my sincere gratitude to everyone who helped me over the years, but there are a few that I need to name personally.

To my committee members, you have taught me, challenged me, corrected me, and made me a better scholar. I could not be more grateful.

Dr. Cheek, I will be forever thankful for the opportunity to work with you in the Postsecondary Education Research Center. What an opportunity to learn about higher ed from one of the best! You have always been eager to share your wisdom and I will be a better administrator and overall human because of my years spent with you. Thank you!

Dr. Zaretzki, you reminded me that we have actually known each other for about a decade. How time flies! You have always been approachable and honest with your feedback. That combination makes for an excellent professor and committee member. I cannot thank you enough.

Dr. Srinivas, though we have never met in person (your face is, like, 2 inches tall in my mind :), you have given to me freely of your expertise in the field of facial expression recognition. My research was improved immensely by our weekly meetings early on, and your timely advice in the latter months. I am so appreciative.

And to my committee chair, Dr Biddix. I feel so lucky to have had you as my advisor. You had a goal to incorporate new ideas into the Postsecondary Education Research Center, and I have benefited from your vision. I got an opportunity to combine two of my passions and to craft an interdisciplinary education in both higher ed and data science. I have learned so much from you these past few years. Your dedication to your family, to your students, and to your research has truly been inspirational to me. Thank you for modeling for me what a well-rounded scholar looks like. I wont forget.

To my church family, I am not exaggerating when I say your presence in my life has made all the difference. I have felt bolstered by your prayers, inspired by your words of encouragement, relieved by your offers to babysit, delighted by your phone messages and funny texts, and strengthened by your unwavering belief that I could finish. Thank you.

To my Dad, Richard, and my siblings, Monica, Cyndela, April, Richard Jr, Darrell, Dimitra, and LaEla and their families; I have felt a new kind of gratitude for you these past few months. Thank you for letting me bounce my ideas off of you, for letting me practice with you, and for offering encouragement and constructive criticism. I needed all of it. You are all precious gems.

To my kids, Myah and Kory, thank you for giving me space and free time to finish this. Thank you for insisting on snuggles, hugs and short talks when you absolutely could not wait any longer. Thank you for so much grace in allowing your mom to be a student. And thank you for being patient. You are the best.

To James, my love. It probably goes without saying I would have never made it without you, but I'll say it anyway. I would have never. ever. made it. without you. Thank you so much for believing that I could do this, even when I forgot. Thank you for being willing to shoulder extra responsibility in our family so I could finish. Thank you for keeping me focused and on task. Thank you for being a rock solid partner. Thank you for being you.

Glossary

Affective Tutoring System (ATS) Educational technology or software designed to incorporate elements of affective computing and emotional intelligence into the tutoring process. These systems aim to enhance the learning experience by recognizing and responding to the emotional and affective states of the learner. 3, 25

arousal The intensity or activation level of an emotion. It indicates how energized or calm an individual feels in response to a stimulus. Emotions can be low in arousal (e.g., calm, relaxed) or high in arousal (e.g., excited, anxious). 6, 39, 66, 81

Cohen's d calculation An effect size statistic used to indicate a difference in outcomes between an experimental group and a control group. A standardized difference between the two means.

$$\frac{\text{mean difference}}{\text{standard deviation}} \tag{1}$$

Cohen's d can range from 0 to infinity. The greater the number, the larger the effect. In general, a calculation from .2 to .49 indicates a small effect, between .5 to .79 indicates medium effect, and .8 or greater indicates a large effect. vii, 15

confusion matrix a cross-tabular matrix table which highlights the performance of a classification model on a given dataset. It compares the ground truth classifications to the predicted classifications to determine the accuracy of a model. Also called an error matrix. 9, 62

convolutional neural network (CNN) A machine learning technique at the heart of deep learning. Used often in image processing, it uses specialized computer algorithms to identify key features and classify images. 3, 26, 47, 94

Cramer's v calculation An effect size statistic used to show the strength of association between two categorical variables.

$$\sqrt{\frac{\chi^2}{N(k-1)}} \tag{2}$$

where χ^2 = Chi square statistic, N = sample size, and k = the lesser number of categories. Cramer's v can range from 0 to 1. The greater the number, the larger the effect. However, a categorization of small, medium and large effect size depends on the degrees of freedom. vii, 17

deep learning A type of machine learning that employs convolutional neural networks to extract increasingly abstract features from data. 3, 26, 31, 57, 81

effect size a statistic that compares two research outcomes and explains the practical significance of the difference between the two. See Cohen's d calculation, Cramer's v calculation, or Hedge's g calculation for more detailed information on specific effect sizes. 15

Ensembles with Shared Representation (ESR) A FER model architecture where a single convolutional network evaluates an image and eventually branches into several distinct networks. Each branch has weighted abstract features slightly differently, producing several predictions. These predictions are combined (in this study by simple voting) to produce a single prediction for the ensemble model. 6

epoch When training a machine learning model, an epoch is one complete pass through the entire training dataset. During each epoch, the model computes the loss function, adjusts its scalar weights, and updates its parameters to improve performance during the next epoch. 9, 31, 60, 90

facial expression recognition (FER) Technology or software used to automatically identify human emotions based on facial expressions. 3, 24, 30, 93

ground truth The objectively correct information about a dataset. This study begins with the assumption that tutors are the knowledge keepers about how to assess student emotions when they are being tutored. 4, 28, 34, 57, 82

Hedge's g calculation An effect size statistic used to indicate the difference in outcomes between an experimental group and a control group. Usually used instead of Cohen's d when the sample size is small.

$$\frac{\text{mean difference}}{\text{standard deviation}_{pooled}} \tag{3}$$

where

$$\text{standard deviation}_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}} \tag{4}$$

Hedge's g can range from 0 to infinity. The greater the number, the larger the effect. In general, a calculation from .2 to .49 indicates a small effect, between .5 to .79 indicates medium effect, and .8 or greater indicates a large effect. vii, 22

in the wild A description of the performance of convolutional neural networks in real-world, uncontrolled, and diverse environments or a description of the datasets these networks are trained on to improve said performance. 10, 84

Intelligent Tutoring System (ITS) A computer-based educational technology that provides personalized and adaptive instruction to learners. ITS is designed to emulate the role of a human tutor by assessing the individual needs and abilities of the learner and delivering customized instruction. 1, 21, 31, 83

JMP Statistical software developed by SAS Institute. All statistical analyses in this study were done with JMP. See the complete output from JMP in Appendix C. 52, 67

RMSE Root Mean Squared Error. A measure of differences between predicted values and observed values in a dataset. Used to assess the accuracy of predictive models (see section 3.9.1). 67, 83

tutoring The act of providing supplemental instruction to a student in specific subject on a smaller scale. For the purposes of this research, scale was 1-on-1 tutoring, and the specific subject was Business Calculus. 1, 15, 30, 57, 81

valence The positive or negative quality of an emotion. Emotions can be categorized on a spectrum ranging from positive (e.g., joy, happiness) to negative (e.g., sadness, anger). Valence helps describe whether an emotion is pleasant or unpleasant. 6, 39, 66, 81

Abstract

When it comes to tutoring, computers have not quite been able to achieve the success that humans have in helping students improve learning outcomes. This research sought to address one aspect of what makes human tutors more effective, the ability to identify and to interpret facial expressions. When a student is feeling anxious, confused, distracted or frustrated, or when a student has an 'aha' moment, human tutors can identify the student's facial expressions and adjust their tutoring approach as necessary. This study sought to determine if, in the context of a gateway college math course, these particular learning-centered affects could be accurately identified by a facial expression recognition deep learning convolutional neural network. Two models were tested. The first model was trained on ground truth images collected during tutoring sessions and annotated by the tutors themselves. This model achieved an accuracy of .25, failing to classify most facial expressions. The second model attempted to predict facial expressions using valence/ arousal matrix scores. This model was trained on Affectnet and achieved RMSE scores consistent with other models trained on the same database. A closer look at the data revealed that although RMSE scores were comparable, there were large areas of facial expressions where the model misidentified the valence in particular. This result suggests that while the valence/ arousal matrix was a better fit in terms of identifying learning-centered affects during tutoring, using the large emotion recognition image databases will not always produce the desired interpretation.

This indicates the need for additional exploration into training valence/arousal models using images specific to tutoring scenarios.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Using computers as an aid for effective tutoring and instruction has been a long-held goal for educationally-minded computer scientists for more than half a century. The ultimate goal is to create a wholly computer-based tutor that is as effective as a human tutor in helping students master a particular subject. The first computer-aided instruction systems of the 1950s simply relayed information, then quizzed the student at regular intervals to ensure comprehension (Morales-Rodríguez et al., 2012). With the advent of artificial intelligence and machine learning, computer tutoring has taken a more adaptive approach, attempting to adjust content and tutoring techniques according to the needs of the student (Anderson et al., 1985). Scientists and researchers seek to craft computerized tutoring systems that will not only detect incorrect answers to questions, but also can diagnose the source of the student's misunderstandings. This allows the system to apply detailed remedies to the specific area of need, and thus become more effective. This transition is marked by the migration from computer-aided instruction to Intelligent Tutoring Systems (ITS).

In the 1990s, ITS seemed poised to be the next great technology in learning analytics; however, the realization of that early promise has been difficult to realize. Until recently, machines powerful enough to calculate adequate responses during student interaction have not been cost effective. Advances in machine learning in the past decade have reinvigorated the pursuit of the most useful ITS. Deep learning, a subset of machine learning, involves building computational models with multiple layers of processing that approach data in increasingly abstract ways (LeCun et al., 2015). This multilayered approach has led to dramatic improvements in classification techniques which are used in many research areas, including visual recognition. This resurgence has made the prospect of useful ITSs viable once again.

Researchers have begun incorporating facial expression recognition software into ITSs. These deep learning classification tools may help ITSs better perceive student comprehension by identifying key expressions during tutoring sessions and using that information to assist the system in identifying student needs. The goal of this research is to add to this body of knowledge and practice by applying two different facial expression recognition algorithms to videos of students being tutored by human tutors, and comparing the results. If facial expression recognition programs can accurately classify facial expressions of students who are being tutored, then that knowledge can be used to make ITS more sensitive to tutee affect and so provide proper treatments to improve student outcomes.

1.2 Statement of the Problem

Research results regarding the efficacy of ITS are mixed. While a few studies indicate that ITS efficacy meets or exceeds human tutoring (Kulik & Fletcher, 2016), most researchers have concluded that ITS are minimally-to-moderately effective in improving student success in

core subjects. In an analysis of 19 controlled studies which included approximately 10,000 K-12 students, Xu et al. (2019) determined that ITS had a small to medium effect on improving students' reading comprehension when compared to human tutoring. A meta-analysis of studies when reviewing mathematical learning among the same age group produced similar results (Steenbergen-Hu & Cooper, 2014).

Strategies for improving efficacy are numerous. There has long been interest in capturing students' affective state as input for ITS to help manage the learning module portion of an ITS. In the late 1990s researchers attempted to capture students' affective state, coining the term Affective Tutoring System (ATS) and incorporating the observation of student affect into ITS research. Observation of speech, gestures, facial expressions, and heart rate were used to attempt capturing student affect (Picard, 2010). Facial expression recognition was often performed using the latest machine learning algorithms, including logistic regression, decision trees, and support vector machines. However, most algorithms were not powerful enough to properly identify and categorize facial expressions. Additionally, initial labeling of the facial expressions was often generic in nature. This may not be the best approach when dealing with the subtle expression changes that are often observed during tutoring.

As artificial intelligence has become more sophisticated, deep learning has replaced other machine learning algorithms in facial expression recognition technology. Deep learning employs multi-layer convolutional neural network (CNN) models to perform increasingly abstract feature extraction to be used for classification. While this technology has become popular in business and health industries, its use in education remains relatively small. The natural progression to using deep learning for facial expression recognition (FER) software may help improve the success of identifying and classifying facial expressions for use in ITS. A marked improvement in this area could greatly improve the efficacy of ITS.

One notable challenge is that ITS research has focused mainly on primary and secondary education. There is little recent literature on the use of ITSs in higher education, and even

less found on facial expression recognition software being combined with ITS to assist in tutoring students in college subjects. If ITSs are going to make a difference for college students, there are nuances that make a focus on higher education necessary. In higher education, the subject matter is more complex, the students may have different motivations and different pressures that may change interactions, and assessment may vary considerably. The research goal of this dissertation is to initiate filling these gaps to improve ITS for higher education.

The logical first step in using deep learning in ITS research is to determine whether the facial expressions of college students who are being tutored in a gateway math course (specifically, basic calculus) can be correctly identified by a FER model in the same way human tutors would. Using human tutors as the ground truth for classifying these expressions is key. If ITSs can ever become as adaptive as human tutors in meeting students' needs, then methods of assessing student comprehension need to improve. Implementing FER models to assist in identifying when a student is confused, frustrated, anxious or distracted can help ITSs develop nuanced responses for each affective state.

1.3 Design Overview

This deep learning experiment is designed to evaluate the effectiveness of two different algorithms designed to classify or identify facial expressions. Both models use a convolutional neural network architecture to identify properties within the tutoring videos. These properties, or features, begin as easily identifiable characteristics on a face (eyes, mouth, etc), but the features become increasingly more complex and abstract as more layers are added to the model.

The experiment began with the collection of video recordings of students being tutored by a human tutor in basic calculus. Approximately 11 hours of tutoring (15 sessions) were

recorded in the Spring of 2022. Each video was reviewed and instances of 5 different facial expressions were identified: 'aha', 'anxious', 'confused', 'distracted', 'frustrated.' These categories were chosen based on research by Lehman et al. (2008), which identified 4 different affective states that influenced learning. The goal of labeling these states was to identify when a tutor might have believed it was appropriate to engage the tutee in a different approach to learning so as to increase understanding of the subject.

Humans can often diagnose complex emotions in a single glance, and it stands to reason that this ability is also used during a tutoring session. Three of these states (anxiousness, confusion, frustration) are indications that a different approach may benefit the tutee at that moment. One state, 'aha', may indicate that the tutee understands the concepts and is ready to move on to the next phase of tutoring. 'Distracted' is an added classification that indicates a low engagement level, regardless of emotional state. 'Unclassified' is a catch all category that encompasses facial expressions not in previous categories.

Tutors, as the subject matter experts, assisted in labeling the recordings. It is often unclear in previous research who was involved in coding the ground truth data used to train the models. When it was mentioned, usually a researcher completed the task (Alexander et al., 2008; Lehman et al., 2008). Since my goal was to have the neural networks mimic the classification decisions of the tutors themselves, it was imperative that tutor intuition drive the coding of the ground truth.

1.3.1 Model 1: Wide Ensemble-Based Convolutional Neural Network

This model approach began with modifying deep learning code originally developed at the University of Hamburg (Siqueira et al., 2020). This method evaluated every nth still frame in a given video. A half-hour tutoring session recorded at approximately 30 frames per second,

produced about 54,000 still frames to evaluate. The algorithm gathered the frames, then cropped and resized the image to contain only the tutee's face in a 125x125 pixel square. The model employs Ensembles with Shared Representation (ESR) architecture (Fig 1.1). The ESR model began with a single convolutional layer that evaluated all pixels in a single still picture. Subsequent layers branched and focused on different aspects of the picture to extract more abstract features. Early layers of the ensemble model are shared among the branches, which reduces redundancy and overall training time. This particular network model expanded into 9 branches. Eventually, each branch outputs a classification conclusion. The ensemble classification is the grouping of the branch classifications combined by simple voting.

1.3.2 Model 2: High-Speed Face Emotion Recognition

This model approached the problem of facial expression recognition using a simplified but seemingly accurate training procedure. The researchers (Savchenko, 2022) developed a neural network package designed to accomplish several related tasks, including facial expression recognition. The result produced a series of related lightweight CNNs, each narrowly trained on a singular task. My dissertation study used one of these networks, *enet_b0_8_va_mtl*, which classifies facial expressions by predicting valence and arousal (V/A) scores (Fig 1.2).

1.3.3 Evaluation of Models

After all videos were coded, they were split into training, validation, and test datasets to provide an evaluation of the fit of the final models. The training data and validation datasets were used to train the ESR9 model. Retraining the model entailed the same training and validation techniques the original researchers first employed, including 10-fold cross-

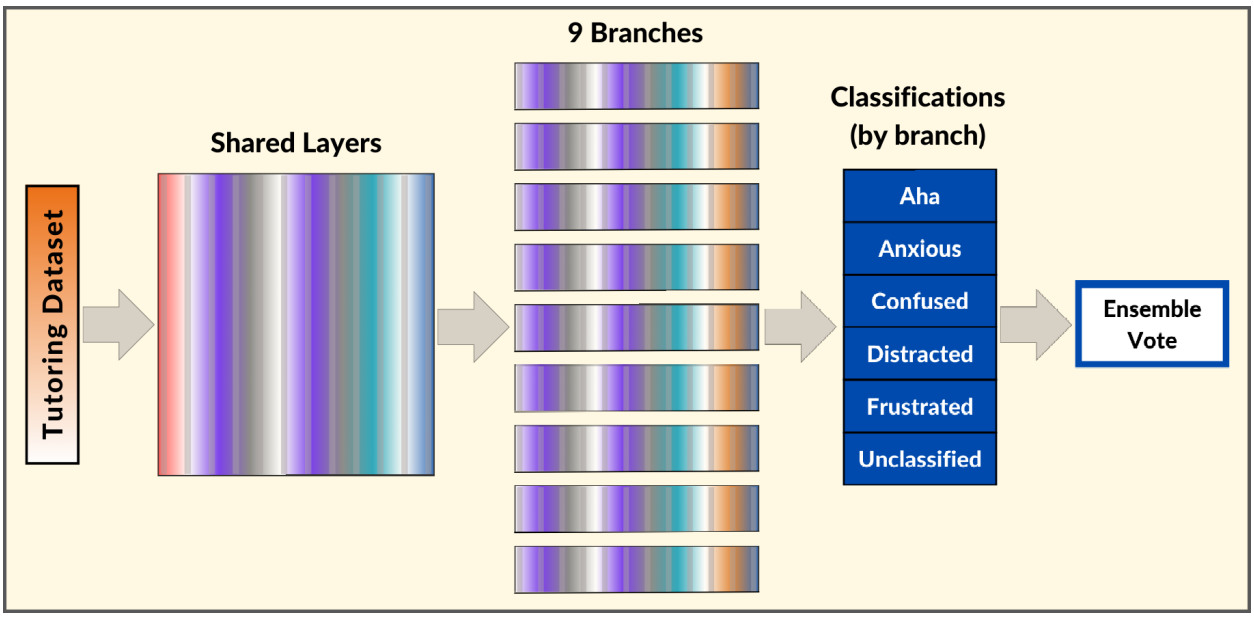


Figure 1.1: Model 1: ESR9 Convolutional Neural Network Architecture.

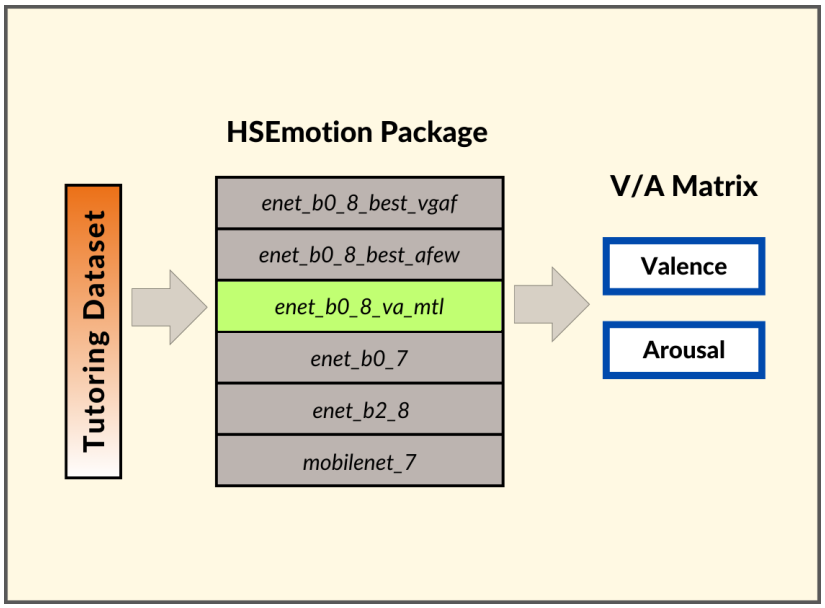


Figure 1.2: Model 2: HSEmotion tool.

validation, and training the first branch with 50 epochs. The HSEmotion model was trained using more than 500,000 images from the AffectNet (Mollahosseini et al., 2019). These images have V/A scores ascribed to them by trained annotators from the University of Denver. The model is trained in only 10 epochs, making it much more lightweight than most other facial expression recognition models, capable of being used in smaller mobile devices.

The student test data was submitted to both models to see how accurately each model could identify and properly classify facial expressions according to the ground truth. Because the ESR9 model was originally trained to classify images in categories typically used in FER research, the model needed to be retrained based on the classifications needed for this study. The architecture of the model was not changed, save for the modifications necessary to accommodate the updated classifications. The model was retrained using python programs provided by the original researcher and modified for this study. Model training was accomplished by adjusting scalar weights to minimize the loss function while increasing the prediction accuracy. Since the model had multiple classification choices, a confusion matrix was used to compare the prediction classes with the ground truth to produce accuracy, precision, and F1 scores (see Section 3.9).

The HSEmotion model calculated V/A scores. Since these values are independent of classification, retraining the model was unnecessary. The HSEmotion model was evaluated by comparing the V/A scores assigned to the student dataset based on the tutor's classification with the V/A scores calculated by the model.

1.4 Statement of Purpose

The purpose of this study was to determine whether the facial expressions of college students who are being tutored by a human in basic calculus could be accurately identified and classified by a facial expression recognition model.

1.5 Research Questions

This study answered the following questions:

RQ1. Can certain facial expressions of higher education students being tutored in basic calculus be accurately classified by a facial expression recognition deep learning model?

Gaining answers to this main question was the primary purpose of this study. The facial expressions of interest are 'aha', 'anxious', 'confused', 'distracted', and 'frustrated'. Identification was considered accurate if the test data were submitted to the CNN model and the model chose the same classification as the tutors. If the results indicated a significant level of accuracy, then an inference could be made that a FER model could successfully mimic a human tutor in terms of identifying key facial expressions.

RQ2. Can models trained on large publicly available datasets be effectively used to assess students' emotional states during basic calculus tutoring?

Skipping classification and instead using V/A scores is an alternate way to identify facial expressions. The emotions of interest certainly span all forms of human experiences. But it is unclear if the facial expressions representing these emotions translate from one experience to another. For example, is the confused expression of a student being tutored in basic calculus sufficiently similar to the confused expression of a person trying to understand the lyrics of a song? The HSEmotion model calculations serve as a proxy for the broader population. Trained on the extensive AffectNet database, which boasts over a million facial expressions, this model is a robust indicator of the broader facial expressions found within the general population. If the V/A scores predicted by the model closely aligned to the V/A scores assigned to the test data by the researcher, then we can infer that existing databases of facial expressions in the wild can be used to train future ITS models.

To evaluate the relationship between the ground truth V/A scores and the scores computed by the model, we investigated a correlative relationship and considered the mean scores. This study began with two null hypotheses:

$H1_0$. There is no linear relationship between valence or arousal when comparing tutor defined V/A scores to the model calculated V/A scores.

$H2_0$. There is no difference in mean valence or mean arousal scores when comparing tutor defined V/A scores to the model calculated V/A scores.

RQ3. Do facial expressions alone offer enough information for convolutional neural networks to accurately assess tutee comprehension?

This question represents a secondary line of research. While the FER algorithms used in this research only have the tutee's face as input, the tutors and researchers who classified the ground truth had the benefit of the entire video to accomplish this task. Additionally, even the accurate classification of facial expressions may not be enough information for an ITS to assess student comprehension. If accurate classification is possible, then answering this research question may help inform and determine future research in this field.

1.6 Significance of the Study

Though ITSs seemed to have a promising future in the late 20^{th} century, improving their efficacy to match human tutoring has proven elusive. Researchers in the 1990s suggested that monitoring the affective state of tutees could be beneficial to improving ITS efficacy (Klein et al., 1999). Since then, methods to identify and monitor affective states have taken a drastic step forward with the advent of artificial intelligence, specifically deep learning. This research has the potential to advance the field of intelligent tutoring by communicating

students' affective states in a timeframe that allows ITSs to adjust their tutoring strategies to meet the needs of the students in real time.

Continued improvements in ITS technology could greatly benefit students enrolled in college math courses. Students with remedial math requirements have the lowest persistence rate of all college students (Kosovich et al., 2019). Due to high failure rates, math classes often become 'gatekeeper' courses, preventing students from continuing in their intended major. Higher education institutions across the country are working hard to change these statistics. But even with the addition of success centers, academic coaches, and extensive tutoring schedules, many students are still failing. While it is unlikely that intelligent tutors will soon (if ever) take the place of human tutors, there is a critical and growing need for expanded student academic support systems in higher education. Improving college math ITS could aid institutions in reducing the failure rate of these core math courses and aid students in reaching graduation.

1.7 Limitations

Detecting emotions is an extremely complex task. Even humans display a wide variance in accuracy when it comes to correctly identifying the emotions of others. Although facial expressions are a primary method humans use to communicate feelings, they are not the only way. One limitation of this study is that there are many other ways tutees may communicate a lack of understanding such as through body language, vocal distress, written cues and other non-verbal communication strategies. These methods are not being evaluated in this study. Further research would need to include other methods of emotive communication for ITSs to become as efficient as possible in detecting the most nuanced of student emotions.

Additionally, the scope of this research on facial expressions is mainly focused on those expressions that indicate a negative valence. This focus was intentional, given that the goal

is to aid ITS in identifying when a new approach is needed to aid in student comprehension. However, students may express a full range of emotions while being tutored, both positive and negative. A broader study of facial expressions in future research would likely benefit the work of building a more affect-responsive ITS.

Another limitation of note is the variation in facial expressions and facial expression detection based on cultural and racial differences. While Darwin proposed that the basic facial expressions were sufficiently similar across cultures (Darwin & Prodger, 1998), researchers through the years have questioned this conclusion (Ekman, 1971). Current research indicates that though there is some commonality among facial expressions, cross cultural variations do exist and are an often overlooked factor in FER models (Sohail et al., 2022).

Finally, the necessary limitation of the population being observed (9 students enrolled at 1 institution during 1 semester) limits the possibility of the conclusions being broadly generalizable. Though the number of hours of video obtained is sufficient for proper training, validation and testing phases, this research more accurately represents an initial proof of concept. If tutee expressions can be correctly identified with these limitations, it may be beneficial to expand the study to include more students.

1.8 Organization of the Study

Chapter 1 identified the research problem, provided a brief overview of the research design, and introduced the research questions this study will address. Chapter 2 provides an overview on the history of tutoring and highlights the most recent research on using artificial intelligence for facial expression recognition in the context of intelligent tutoring systems. Chapter 3 outlines the methodology of this study and provides detailed information on the participants, the data collection, and the machine learning models used to analyze the data.

Chapter 4 provides the results of the study for each model. Finally, Chapter 5 addresses the research questions and provides recommendations for further research.

Chapter 2

Literature Review

2.1 Introduction

Tutoring is a time-tested method of assisting students who are having trouble mastering academic subjects. Tutoring is defined as supplemental instruction given to students in a one-on-one or small group setting in a specific academic area. This instruction is given by an individual (a tutor) who has shown mastery in that subject area (Dang & Rogers, 2008). Tutors typically offer more personalized instruction and so can assist in explaining concepts the tutee may have difficulty understanding.

In this chapter I will first examine the impact of tutoring on student success. Though most research focuses on students at the primary and secondary levels, the limited research at the postsecondary level shows a net positive effect of tutoring in higher education. Effect size is the most used measure to denote the effectiveness of a tutoring method. Note that unless otherwise indicated, effect size is Cohen's d calculation. Next I will examine the use of computers both as a tutorial aid and as the primary tutor. Current research is varied, but most conclude that though computers are helpful as a tutorial aid, students who use computers as their primary source of tutoring do not benefit as much as students who received

tutoring from a human expert. Finally, I will examine the emergence of artificial intelligence and its impact on tutoring.

2.1.1 The Impact of Tutoring on Student Success

Tutoring programs have become increasingly popular in recent decades. In the l990s an estimated 4-6 percent of secondary students received private one-on-one tutoring (Dang & Rogers, 2008). Current estimates are that the rate has grown to about 10 percent of students. Additionally, legislation and policy are already in place to increase that statistic (Sparks, 2023). As higher education has shifted to a more outcomes-based model (Jones et al., 2002), tutoring programs have become commonplace on college campuses. Tutoring is seen as a cost effective way to improve student success.

Despite broad agreement on the efficacy of quality tutoring, there is surprisingly little comprehensive research on tutoring in postsecondary environments. Martha Maxwell, a respected pioneer in postsecondary learning assistance and developmental education, theorized that research regarding tutoring has suffered due to negative stereotypes. Tutoring is often associated with failing grades, poor study skills, or lack of intelligence. Conversely, others believe access to tutoring implies excessive wealth or special treatment (Maxwell, 1990). In her opinion, these biases have resulted in poor funding and poor participation in tutoring research over the years.

Historical research has concluded that successful tutoring is dependent on several variables. These factors include characteristics of the student learner such as motivation, confidence, and cognitive ability, characteristics of the tutor such as preparation, experience, and expectations, and aspects of the student's non-academic environment such as family situation, cultural background, and socioeconomic status (Hartman, 1990; K. J. Topping, 2013). Of these variables, tutor characteristics seem to have the greatest impact on outcomes

for students with the greatest need. Students who are at risk of not completing college due to academic unpreparedness need tutors who are carefully trained. These tutors need guidance on how to work with students with academic deficiencies before they begin tutoring. Cohen's review of tutoring literature found that the largest effect sizes in tutoring were observed in well defined and structured tutoring programs, and in tutoring sessions of shorter duration (Cohen et al., 1982). Additionally, tutors need to be evaluated and receive feedback on a regular basis (Hartman, 1990; K. J. Topping, 2013).

Most tutoring research at the postsecondary level focuses on outcomes for specific populations (STEM students, underrepresented minorities, first generation students, etc. . .). Outcome is measured by test scores, grade distribution, retention, or persistence to graduation. A literature review by Topping indicated that supplemental tutoring for medical students in a given set of courses had an average effect size of between one third and one half of a standard deviation on their test scores (J. K. Topping, 1996). Ticknor evaluated the outcomes of STEM students at his institution who visited the Math & Science Learning Center for tutoring versus those who did not. He reported a small positive effect size (Cramer's v calculation) of 0.09 when comparing the final grade distribution of given courses (Ticknor et al., 2014). A stratified examination of this data also indicated that Black males were even more impacted by tutoring with a medium effect size (Cramer's v) of 0.21.

More recent research has confirmed that specific and individualized instruction improves not only student academic achievement, but also communicative and collaborative skills (Seo & Kim, 2019). Most of the research into tutoring in the past five years has focused on specific types of tutoring programs and their outcomes. Peer tutoring programs in various academic settings were instrumental in improving overall academic performance (Gamlath, 2022). KC et al. (2020) examined pre and post indicators of 28 medical students enrolled in Anatomy and Molecular Medicine courses. They compared their first exam scores (before they received tutoring) with their final exam scores and found a significant difference in

the scores (63.4 vs 69.9 for Anatomy and 65.3 vs 74.4 for Molecular Medicine). Arco-Tirado et al. (2020) conducted a controlled experiment with 102 freshmen enrolled at his institution. Half of the students engaged in weekly tutoring sessions for their first year while the other half functioned as a control group. Their results indicated that the experimental group showed significant improvement over the control group when comparing success rate (credits passed/credits completed) and dropout rates (89% vs 74% and 2% vs 6% respectively). An additional benefit often observed was that peer tutoring proved to be beneficial to both the tutor and the tutee (Arco-Tirado et al., 2020; KC et al., 2020).

A more recent tutoring practice is embedded tutoring. Embedded tutoring is a program where the tutor attends class sessions of the given subject with the tutee. The purpose is to increase student engagement in the primary learning environment and so to improve student comprehension. Though the time commitment for tutors is significant, outcomes included improved individual skills, better course grades and higher student retention. Channing and Okada (2020) compared course sections at an institution that employed embedded tutoring with sections of the same subject taught by the same instructor that did not have embedded tutoring. He found that retention rates were higher for the sections with embedded tutoring (76% vs 71% for English, 52% vs 48% for Math, and 86% vs 82% for Chemistry). Miller (2020) investigated the changes in perception of students enrolled in 3 sections of an engineering course. She determined that students who were involved in the embedded tutoring intervention improved their 'mindset score' regarding writing from a pre-semester score of 4.36 to a post-semester score of 4.71. The control group's mean actually decreased from a pre-semester score of 4.56 to a post-semester score of 4.48. Additionally she found that the experimental groups' writing performance had a significant increase in score between first and final drafts of a writing assignment while the control group did not.

High-dosage tutoring has seen a resurgence in popularity, especially since the COVID-19 restrictions caused such a disruption in primary and secondary learning. Even prior to

the pandemic, first-time freshmen preparedness for college math courses was demonstrably inadequate (Corbishley & Truxaw, 2010; Jaafar et al., 2016; Nihan Er, 2018). The effects of the pandemic have only made it worse. Nationally, students 1st through 8^{th} grade are testing below grade level in math at a higher rate than prior to the pandemic (Moore & Hayes, 2023). Students indicated they felt less prepared for college level math due to restrictions during the pandemic. In a survey conducted by ACT, 57% of entering freshman respondents felt that closures in Spring 2020 negatively impacted their academic preparedness in math for the 2020-21 academic year (American College Testing, 2021). High-dosage tutoring programs seek to mitigate the effects of unpreparedness. These tutoring programs are focused mainly on K-12 education; however, one common goal is increased academic preparedness at the postsecondary level. There is no standard definition of high-dosage tutoring. The National Bureau of Economic Research defines it as "being tutored in groups of 6 or fewer for more than three days per week or being tutored at a rate that would equate to 50 hours or more over a 36-week period." (Fryer, 2016, p.38). deRee et al. (2023) instituted an experiment on high-dosage tutoring in math in 3 primary schools. After 1 year of high-dosage tutoring in math, researchers noted a .28 effect size, which indicates an intervention's effect is large. All of the aforementioned tutoring programs have been shown to improve success rates of postsecondary students. It remains to be seen whether high-dosage tutoring will aid in reversing the trend of postsecondary unpreparedness of students since the shutdown.

A commonly held reference point for tutoring researchers is that, relative to classroom instruction alone, tutoring will impact average student achievement by an effect size of 2. The origins of this standard are a 1984 study (Bloom, 1984). This research indicated that of three different learning conditions evaluated, tutoring was far and away the most effective. But since it is cost prohibitive to hire effective tutors for each student, the 'two sigma problem,' as Bloom labeled it, remains. How can educators help students reach their highest

potential of learning using a more practical approach? This is perhaps the origin of modern day research into computer aided tutoring.

2.2 History of Tutoring Using Computers

2.2.1 Computer Aided Instructional Tools (CAI)

Though the history of the 'teaching machine' goes back at least two centuries, use of modern (digital) computers to improve student learning traces back to the 1950s (Benjamin, 1988). One of the more successful CAI tools was PLATO (Programmed Logic for Automatic Teaching Operations). Introduced in 1960, PLATO was an instructional system developed by researchers at the University of Illinois (Alpert & Bitzer, 1969). This system instructed users in various subjects at primary, secondary, college, and workforce levels. Though not especially profitable, the PLATO platform was in use through the turn of the millennia and spawned several innovations still in use today in online communities (Woolley, 2016).

Likely, one of the more important contributions of PLATO to computer tutoring is the confirmation that CAI tools could indeed be used to improve student learning outcomes. PLATO published independent research results from 49 different studies (Foshay, 2004). Of these, results were reported for 997 users at 7 postsecondary institutions. In one study, a math department at a community college conducted an experiment with 35 students in the experimental group and 46 students in the control group. A math pre-test determined the two groups were statistically identical. The post-test results revealed that the experimental group who used the full implementation of PLATO showed a 24.4% gain while the control group showed a 14.4% gain. The outcomes from these multiple studies were not standardized, so it is difficult to compare the results. However, each study indicated that on average, college

students who used a full implementation of the platform consistently performed better on standardized and published tests than students who received no tutoring.

The ubiquitous influence of computers in our world today means it is nearly impossible to study the effects of tutoring without some sort of computer-aided instruction. Traditional, formal computer-aided supplemental instruction has evolved into game-based learning in primary and secondary education. The latest research indicates mixed results. Ito et al. (2021) conducted an experiment in Phnom Penh, Cambodia to research causal impact of CAI-based software on primary student outcomes in a developing country. The experiment group used a CAI game called Think!Think! for 30 minutes each weekday for 3 months. Comparison of math achievement tests between the experimental group and the control group indicated a positive effect size of .68-.70. Conversely, researchers in India reported different results from a similar experiment. deBarros and Ganimian (2023) focused on whether primary students who were lagging behind curricular expectations could improve their math performance by using computer aided instructional tools to 'bridge the gap.' After 4 months, they reported that computer aided instruction had a statistically insignificant improvement in math achievement of the average student in the sample. Relative to the control group, the treatment group effect size was .05, well within the standard error range. Furthermore, their results showed no improvement when treatment was increased to 9 months. These two experiments are not identical, but do highlight the range of results when reviewing the latest research with computer aided instructional tools.

2.2.2 Intelligent Tutoring Systems (ITS)

The term Intelligent Tutoring System (ITS) was introduced by Sleeman and Brown upon publication of their book with this title (Sleeman & Brown, 1982). ITSs are computer systems employing artificial intelligence, cognitive science and educational psychology to

produce platforms that attempt to mimic some aspects of tutoring behavior that humans naturally perform (Nwana, 1990). ITSs are fundamentally more advanced than their CAI predecessors in that ITSs are designed to be highly adaptable based on user needs. Systems might generate questions on the fly that are tailored to the student's abilities (particularly useful in math tutoring), or might adjust tutoring tactics based on the student's learning style (Milne et al., 1997).

There is debate regarding just how effective ITS are compared to human tutoring. Common wisdom from the past quarter century held that ITSs were about half as effective as human tutors in improving learning outcomes. Orey offered a summary of his evaluation of 3 different ITSs. He began his research with the understanding that human tutors have an average effect size of 2.0 when compared to classroom instruction alone. His preliminary expectation was that the ITSs he evaluated should have an effect size of about 1.0 (Orey, 1993). More recently, other researchers have reported more varied conclusions. VanLehn (2011) and VanLehn et al. (2014) conducted meta-analyses of tutoring experiments and found that the best performing ITS were nearly as effective as human tutoring. He determined that far from the standard 2.0 effect size, human tutoring had an effect size of about .79 and ITS had an effect size of about .76. His analysis also concluded that as ITSs increase the number of opportunities students have to interact with the system prior to solving a problem, the more effective the tutoring. A literature review by Feng et al. (2021) highlighted effect sizes (Hedge's g calculation) of .35 to .59 when using an ITS for STEM subjects as compared to no tutoring.

China embraced the ITS concept in education, offering tax breaks and financial incentives for companies that use artificial intelligence to improve student learning (Hao, 2019). Pai performed a 2-week experiment with remedial math students in Taiwan (Pai et al., 2021). 134 students were divided into 3 groups. One group used only a dialogue-based ITS for instruction, another group had conventional teacher instruction, and the third group only

read the materials. The results showed that the first two groups had statistically identical effect sizes (1.09 and 0.95 respectively) when comparing students' pre-tests and post-tests. The researchers concluded that student performance was comparable between the two groups, and that ITSs were just as effective in this controlled setting as a conventional teacher.

However, other research has shown ITS to be less beneficial than what was previously believed (Steenbergen-Hu & Cooper, 2014). Educators in Sweden, for example, experienced unexpected and difficult challenges using an ITS-assisted pedagogy that was designed to complement their instruction. Ultimately, the instructors felt compelled to abandon the confines of the ITS in order to effectively teach their students (Modén et al., 2021).

While the results are varied, it is clear from the literature that students typically benefit from any kind of tutoring. In general, researchers agree that CAI tools benefit students the least and human tutoring benefits students the most (VanLehn, 2011). ITS efficacy falls somewhere in between.

2.2.3 Emotion Detection and Affect Tutoring Systems (ATS)

In 1872, Darwin proposed that, regardless of culture or background, all humans show six basic emotional states in remarkably similar ways: happiness, anger, surprise, sadness, fear and disgust (Darwin & Prodger, 1998). Though he indicated that body language plays a role in communicating emotion, his research focused primarily on the face. His conclusions were further entrenched by the research of Ekman and Friesen, which concluded that along with Darwin's identified emotional states, a facial expression for contempt could be added to the list of universal facial expressions (Ekman & Friesen, 1986). These conclusions have been a contested topic over the years. Nevertheless, the listed classifications have been the basis of a great deal of facial expression research. Modern research in emotion recognition

often relies on some or all of these facial expressions as classification options (Behera et al., 2020; Bhatti et al., 2021; Yu et al., 2021).

The practice of adding affective state processing to computers to achieve more nuanced human and computer interactions began in the early 2000s. Though users regularly communicate emotions while using computers, the computers were not equipped to identify or utilize this important data. Turn of the century research indicated that even if computers were not sophisticated enough to fix issues, acknowledging users' emotions with phrases like "we apologize for your inconvenience" were helpful in reducing negative emotions when they were identified (Picard, 2000).

Today, research into facial expression recognition (FER) remains a daunting task. Different facial expressions notwithstanding, changes in head orientation, background, obscured view, and lighting all contribute to the complexity of the problem (Devi & Preetha, 2021). Additionally, humans are likely to use much more than facial expressions to interpret emotional states. However, computer science researchers understand that facial expressions are the best single predictor of intense emotions.

Adding this same affect-processing to ITSs is a natural progression for computer tutoring. Since ITSs are designed to adapt based on user needs, it follows that student affect must be taken into consideration when possible. Longstanding research indicates that a student's affective state influences his or her ability to learn during a tutoring session (D'Mello & Graesser, 2012). An important skill among expert human tutors, therefore, is the ability to understand the emotions and demeanor of the students they tutor. Understanding a tutee's emotional state gives a human tutor the ability to direct the tutoring session in the most productive manner. Lehman et al. (2008) conducted an experiment and tested for Darwin's 6 different basic emotions, 6 additional learning-centered emotions, and 4 different engagement levels. She found that happiness was the only basic emotion that commonly occurred during

tutoring sessions. Of the learning-centered emotions confusion, anxiousness, and frustration were the significant affective states.

Affective Tutoring System (ATS) often attempt to employ two methods of affect-state interpretation. First is detecting emotions based on physical characteristics such as facial expressions, speech patterns, or autonomic responses such as blood pressure and heart rate. The second method is affect prediction based on students' personality type and previous interactions. Regarding physical characteristics, research projects from the 2000s classified facial expressions almost exclusively by examining video footage (Alexander et al., 2008; Lehman et al., 2008; Person & Graesser, 2003). In each instance, the classification was completed by researchers who carefully reviewed the video and determined the best category for each facial expression they observed. When feasible, more than one researcher worked together to classify the facial expressions viewed in the videos. But due to the volume of footage to review (even an hour of video may result in hundreds of video clips to be examined), a single researcher was responsible for coding and classifying the majority of facial expressions. The clinical nature of classification and the different categories of classifications resulted in a wide variability among experiments. Alexander et al. (2008) coded nine different facial expressions and also recorded an intensity of high or low. Lehman et al. (2008) coded 12 different facial expressions with 4 different engagement levels. Like emotions themselves, affect classification remains varietal and difficult to label.

2.3 Using AI for Facial Expression Recognition in Intelligent Tutoring Systems

Adding FER programming to ITS to interpret tutees' affect is not a new concept, but early attempts were either unsuccessful or infeasible. The earliest research explored facial

expressions as a method of communicating to, rather than receiving input from, the tutee. Towns et al. (1998) developed an entire emotive behavior framework designed to replicate human emotive responses. This framework was developed into a lifelike (stylized) tutor named Cosmo that instructed students about the Internet. Cosmo was an ITS that used facial expressions and body language to communicate causal responses to the student's actions including congratulatory, interrogative and inquisitive speech.

Personal computer systems in the early 2000s did not yet have the computational power to provide useful feedback to an ITS in a live context, but the concept was being explored. Lisetti and Schiano (2000) wrote a multidisciplinary article explaining the need for human-computer interaction scientists, artificial intelligence researchers and cognitive science experts to begin to work together to solve the problem of automatic facial expression interpretation. In 2007, a research team at the University of Memphis added affect recognition to their ITS (named AutoTutor) in an attempt to identify and respond to student emotions. AutoTutor used conversational cues, body posture, and facial tracking to recognize these emotional cues. IBM's BlueEyes technology was incorporated into AutoTutor to achieve facial tracking (D'Mello et al., 2008). BlueEyes was an IBM product (both hardware and software) used to assist computers in sensing human emotions. AutoTutor's facial tracker labeled facial action units with a 68% accuracy. IBM did not publish the proprietary workings of BlueEyes technology, but presumably, they used the most popular machine learning techniques of the time, including Bayesian inference models and support vector machines.

The recent growth of machine learning technology (specifically, deep learning) has impacted the field of computer vision research tremendously. Convolutional neural networks have transformed the landscape of artificial intelligence. Large databases of facial images have been developed to train and test deep learning models in attempts to enhance facial recognition systems, leading to substantial advancements in accuracy and robustness (Kollias

et al., 2019; Mollahosseini et al., 2019). Despite challenges These datasets have facilitated breakthroughs in facial feature analysis, emotion recognition.

With this growth, attention is again turning to facial expression recognition enhanced ITSs. Computer scientists presented research at the 2021 International Conference on Human-Computer Interaction regarding using facial expressions as indicators of engagement when students are using ITS (Yu et al., 2021). Their research used the Facial Action Coding System (Rosenberg & Ekman, 2020) to detect three action units (smile, nose wrinkle, frown) using a deep learning model. This information was communicated to the instructor via a Teacher Dashboard. Future work for this research involves combining the FER with their head pose detection system to ensure the most sensitive method of emotion detection.

2.4 Summary

It is clear from the research that tutoring as a supplemental instructional tool can have a major positive impact on student learning. Unfortunately, the most effective method of tutoring is often cost prohibitive for the students that need it and would benefit from it the most. Educators have been searching for decades for ways to apply the benefits of tutoring in a cost effective manner. Intelligent tutoring systems may prove to be the best option provided ITSs can become more responsive and adaptive as effective human tutors. Technological advancements in the past decade have led to increases in the commercial use of FER models (Dupré et al., 2020). While this has led to a revival in research that focuses on student affect when using an ITS, little research has been done on classifying tutee facial expressions when interacting with a human tutor. This is a vital step in the journey to improve ITS interaction with humans. Current research typically uses the basic FER classifications that are not specific to tutoring, and uses researchers to manually classify training data. However, human tutors have honed skills in interpreting students' facial expressions. They may be

the best source in classifying facial expressions. By using human tutors' classifications as the 'ground truth' or standard for determining what student emotions during tutoring, this research is contributing a vital study to the corpus of ITS research.

Chapter 3

Methodology

3.1 Introduction

This chapter describes the research methodology and elaborates on the design decisions implemented in this exploratory study. The chapter provides a detailed description of the research design, the data collection, the analysis plan, and an overview of validity confirmation. This research sought to answer the following questions:

RQ1. Can certain facial expressions of higher education students being tutored in basic calculus be accurately classified by a facial expression recognition deep learning model?

RQ2. Can models trained on large publicly available datasets be effectively used to assess students' emotional states during basic calculus tutoring?

To evaluate the relationship between the ground truth V/A scores and the scores computed by the model, this study began with two null hypotheses:

$H1_0$. There is no linear relationship between valence or arousal when comparing tutor defined V/A scores to the model calculated V/A scores.

$H2_0$. There is no difference in mean valence or mean arousal scores when comparing tutor defined V/A scores to the model calculated V/A scores.

RQ3. Do facial expressions alone offer enough information for convolutional neural networks to accurately assess tutee comprehension?

3.2 Research Design

Exploratory research often is associated with qualitative studies; however, this approach is well established in machine learning research. My quantitative research study was best framed as exploratory for the following reasons. First, the research questions lent themselves to exploring new opportunities in an established paradigm. Discrete facial expression classification today is almost universally confined to the six basic emotions (happy, sad, surprise, fear, disgust, anger) and sometimes contempt and neutral (Li & Deng, 2022a). Contemporary peer-reviewed research on facial expression recognition (FER) using other categories is scant. Prior research indicates that the emotions expressed in a tutoring session will likely not utilize all six basic emotions (Bosch et al., 2015). In an effort to tailor the research to adequately answer the proposed research questions, other emotion classifications were used. In this case, learning-centered affect states first categorized as such in the early 2000s (Lehman et al., 2008) were again brought to the forefront. Answering the research questions through exploratory methods was used to reveal whether it would be necessary to expand the basic emotion paradigm to include other emotion classifications.

Second, an exploratory research design is flexible, which was needed when working with this image dataset. The data collection process during this study presented several challenges. Although there were thousands of data points per student, the limited number of participants represented in the dataset confined the research corpus of facial expressions to those expressed by a few people in a relatively short amount of time. Conversely, the

amount of footage collected (over 11 hours of footage resulting in over 1 million still frames) categorized it as a large dataset. An image dataset this size has its own difficulties in terms of storage and data processing. Machine learning models require a large volume of data in order to be trained properly, however, practical considerations of time and expense often mean models may go through limited rounds of complete testing (called epochs). When training the discrete model for this study, the wide ensemble CNN titled ESR9 (see Section 1.3.1), the documentation recommended 50 epochs for the shared layers and 20 epochs each for subsequent branches. The issues of limited participants, a very large dataset, and possibly limited testing epochs made it difficult to achieve concrete solutions to the research questions through this study.

Finally, this research represents an initial step in solving a larger problem. As discussed in Chapter 1, the main issue is that Intelligent Tutoring Systems (ITS) are not as effective as human tutors in helping postsecondary students. Conducting this research was not intended to solve this problem. This study was an opportunity to explore one aspect of the issue to determine if a portion of the problem can be solved with deep learning.

3.3 Research Setting

This research data was collected in the Spring semester of 2022. All participants were enrolled in a business calculus course at a large, public research university. The Institute for Higher Education Policy defines a gateway course as a college-level, introductory math or English course (Janice & Voight, 2016). Business calculus was chosen because it is the gateway math course to many majors. It is also considered a "bottleneck" course, meaning that many students have trouble passing the course and so are unable to progress to their major. Pragmatically, this is defined as a course having a high drop, withdrawal or fail (DWF) rate. Institutions recognize this issue and many have set up tutoring centers or

other success centered initiatives to help students in high DWF courses (Jaafar et al., 2016). Employing ITSs that were effective in aiding students in this course would result in higher success rates for both students and institutions.

3.4 Research Participants

The research study included 9 participants enrolled in several sections of a business calculus course. Students were recruited using flyers posted in the common areas of the tutoring center. Additionally, calculus tutors who agreed to participate in the annotation emailed a prepared letter to their regular tutees asking if they would like to participate. Students were offered a $25 Amazon gift card for agreeing to participate. All participants signed a consent form and agreed to be recorded for the purpose of this study. Students could participate up to 3 times, earning up to $75 in gift cards. This study was reviewed and approved by the Institutional Review Board in accordance with FDA regulations (Appendix A).

The original goal was to enroll up to 30 students in Fall of 2021. Each student would participate in at least one 30-minute tutoring session. Unfortunately, the global pandemic and subsequent shift to online class sessions created challenges in achieving student participation. Though many universities had resumed in person instruction by Fall of 2021, much of the tutoring continued to be conducted online at the research institution. Additionally, many of the in-person tutoring sessions were conducted with participants wearing face masks. This is a considerable barrier to facial expression research. Though it was not in the original plan, I determined that postponing recruitment to Spring of 2022 would yield more participants. Additionally including online tutoring into the research pool would be required in order to collect enough footage. Limited demographic information was collected about the tutees (Table 3.1).

Table 3.1: Participant Demographics

Total Participants	9 students
Sex	2 men, 7 women
Race	3 Asian 3 Black\AA 2 White 1 Unknown
Tutoring Location	5 students had 8 tutoring sessions on campus 1 student had 2 tutoring sessions off campus 3 students had 5 online tutoring sessions
Age Grouping	2 students aged 18-20 5 students aged 21-30 3 students 30+
Classification	All students were undergraduates 1 participant was a transfer student
Other	8 students were taking Business Calculus for the first time 1 student was repeating the class

3.5 Tutors

Although tutors were not included in the data collection, they were an integral part of the research. Tutors were provided a $50 Amazon gift card for agreeing to participate and annotating at least one video. They were the leading sources when it came to developing the ground truth for this study. Limited demographic information was collected about the tutors (Table 3.2).

3.6 Data Collection

In person recordings were collected using an iPhone 13 Pro 1080p HD video camera, recording at 30 frames per second (fps). A tripod was set approximately 3-4 feet from the participant, depending on the setup. I attempted to be as unobtrusive as possible in the placement of the camera. The goal was for the participants' behavior to be influenced as little as possible by the camera or presence of any persons associated with this study. Online tutoring was recorded using the application and equipment the participants elected to use for their tutoring session. Online participants used either Microsoft Teams or Zoom to record their tutoring sessions.

Data collection concluded at the end of the Spring 2022 semester with the 9 participants recording a total of 15 tutoring sessions (Table 3.3). As previously stated, I left the tutoring areas (both online and in person) after recording had begun in order to preserve privacy and to minimize the impact of this research on the tutoring session. This meant that as students shifted (e.g. to retrieve a book from their backpack, or to take a drink), they were occasionally out of focus or out of the recording frame. And when the student returned to focus, they might not be in the exact same position. In one online recording session, the student forgot about the recording and turned their camera off a few

Table 3.2: Tutor Demographics

Total Participants	5 tutors
Sex	1 man, 4 women
Age Grouping	2 tutors aged 21-30 1 tutor aged 30+ 1 tutor age unknown
Classification	All tutors were undergraduates 1 participant was a transfer student
Years tutoring	4 tutors had tutored for 2+ years 1 tutor had tutored for less than one year

Table 3.3: Tutoring Sessions

Student	**Recording Number**	**Tutoring Location**	**Recording Length**
Student 1	1	online (MS Teams)	01:17:36
Student 1	2	online (MS Teams)	00:25:59
Student 2	1	online (Zoom)	01:06:14
Student 2	2	in person, on campus	00:49:36
Student 3	1	in person, on campus	00:36:58
Student 4	1	in person, on campus	00:33:49
Student 5	1	in person, on campus	00:36:56
Student 5	2	in person, on campus	00:57:45
Student 5	3	in person, on campus	01:01:21
Student 6	1	in person, on campus	00:56:12
Student 6	2	in person, on campus	00:51:34
Student 7	1	online (Zoom)	00:34:12
Student 8	1	in person, off campus	00:21:45
Student 8	2	in person, off campus	00:31:37
Student 9	1	in person, on campus	00:46:16

minutes into the session to eat. These issues along with the trimming of the videos to only tutoring time, reduced the number of usable hours by about 20 percent.

A total of 11.4 hours of tutoring footage was collected and reduced to about 9.1 usable hours. At a video recording rate of 30 fps, this collection provided an ample amount of images for training, validation, and testing.

3.7 Research Procedures

3.7.1 Discrete Classification of Facial Expressions

The categorization of facial expressions for this study was chosen based on Lehman's research on students' affective states while engaging in tutoring (Lehman et al., 2008). Lehman and team monitored for the 6 basic emotions, 6 different learning-centered affects, and 4 engagement levels. They concluded that less than half of these affective states could be significantly observed during tutoring. They identified 3 learning-centered affects that occurred at significant levels during tutoring sessions (Table 3.4).

Two more categories were added to this foundation. The learning-centered affect 'aha' was included at baseline because, although it did not rise to significance in Lehman et al. (2008) research, I hypothesized that it might be an important facial expression for an ITS to be able to identify if possible. While this study might yield the same results as Lehman et al. in terms of significance, I concluded that it would be best to collect the data and let the model determine significance. An engagement indicator of 'distracted' was added for similar reasons. An improved ITS should be able to gauge the engagement level of students if possible. Finally, 'unclassified' was added as a category to mark all other images in the dataset that were not already categorized.

Table 3.4: Discrete Categorization

Affective State	Definition	Importance
aha	student has a sudden realization about, or understanding of the material.	This affect did not rise to the level of significance in Lehman's research. However it is an important affect in human tutoring and so is worth another evaluation with a new tool. This affect may signal to an ITS to offer an affirmation or that it is time to proceed to the next topic.
anxious	student displays nervousness, anxiety, negative self-efficacy, or embarrassment.	A learning-centered affect that has proven to be significant in previous research. This may be a signal to an ITS to try a different approach to the topic at hand.
confused	poor comprehension of material, leading to attempts to resolve erroneous belief (lower arousal than frustrated).	A learning-centered affect that has proven to be significant in previous research. Though the mean valence score is negative, this affect is often viewed as positive in the tutoring setting. In a confused state, the learner experiences cognitive disequilibrium and is forced to think. The lower arousal score allows the student time to work on the problem, often leading to an 'aha' affective state soon after.
distracted	student is not focused on the tutoring topic.	This may be a signal to the ITS that it is time for a break.
frustrated	poor comprehension of material, leading to exasperation, irritation or annoyance (higher arousal than confused).	A learning-centered affect that has proven to be significant in previous research. This may be a signal to the ITS that it is time for a break.
unclassified	all other facial expressions not categorized above remain unclassified for the purposes of this experiment.	

3.7.2 Continuous Affect

Another favored method of identifying emotions within machine learning models is the valence-arousal matrix (Russel, 1980). Valence and arousal (V/A) are two independent scores which together, map a continuum of emotions. Valence is the degree to which a subject feels the pleasantness of a given moment. Valence score ranges from very unpleasant to very pleasant. Arousal is the degree to which a subject feels emotionally activated or deactivated. Arousal score ranges from low to high. The V/A model suggests that "individuals do not experience, or recognize, emotions as isolated, discrete entities, but that they rather recognize emotions as ambiguous and overlapping experiences" (Posner et al., 2005, p. 4). With this model, emotions are not listed in a discrete category, but rather given a score that can be mapped onto a model of affect. Figure 3.1 depicts a graphical representation of the circumplex model of affect (Posner et al., 2005). This continuous affect method applies a scale of [-1, 1] to both arousal and valence scores. Both models used in this study were initially trained on the AffectNet dataset which is labeled with both discrete classifications and continuous affect (V/A) scores (Mollahosseini et al., 2019).

3.7.3 Annotating the Dataset

Paramount to this research is the tenant that if ITSs are going to improve, they must be able to identify certain human emotions with the same accuracy as the tutors themselves. Therefore, tutors must take the lead in annotating the facial expressions in the research dataset. Of the 5 tutors involved in this research study, 4 tutors annotated the expressions of the students they tutored. One tutor elected not to participate in an annotation session, so another tutor was used to label facial expressions of a student they did not tutor. This resulted in about 70 percent of the footage being labeled by the tutors who conducted the tutoring session and about 15 percent of the footage being labeled by tutors who were not

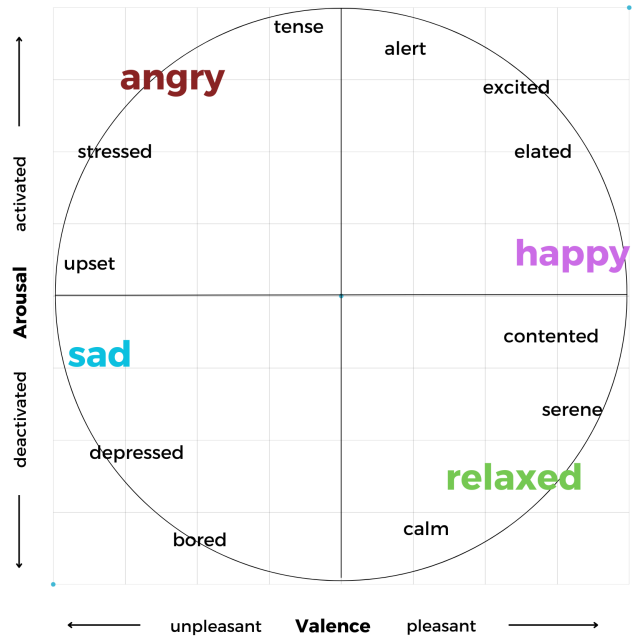


Figure 3.1: A graphical representation of the circumplex model of affect.

involved in the particular tutoring session. The remaining 15 percent of footage was labeled by this researcher after spending several hours with the tutors during annotation sessions. In all, 418 observations were identified and annotated (Table 3.5).

Tutors were instructed to mark the hour, minute, and second of the video when they observed an emotion they believed the student was feeling one of the 5 learning-centered affects. Tutors were not restricted to using just facial expressions during their annotation sessions. They watched and listened to the entire recording. Though not specifically instructed to do so, the tutors presumably used all available information to classify students' emotions. They were able to listen to the dialogue, see the tutee's face and body language, see the tutor's face and body language, and in most instances, recall their own memories at the time of the session. Along with the discrete categorizations, the tutors were also able to indicate a 'more' or 'less' attribute if they believed the student displayed the given emotion 'more than' or 'less than' their typical display of emotion.

Tutors met with me either online or in person to provide their annotative feedback. The tutors indicated the precise time on the video when they believed the student was feeling one of the observed emotions. I manually reviewed the several seconds before and after to capture the complete 'emotion arc.' I then constructed .json files based on the tutors' annotation to indicate frame by frame the classification of the students' face.

Valence and Arousal Scores

The assigning of V/A scores to facial expressions is a tedious and expensive task. The creators of the AffectNet database employed and trained several annotators to accomplish this task for their large dataset (Mollahosseini et al., 2019). For this study, manually assigning accurate V/A scores for the approximately 75,000 images annotated by the tutors was an impossible undertaking. Instead, I elected to programmatically approximate the V/A scores using available information.

Table 3.5: Annotations

Student	Minutes Reviewed	Number of Observations					
		aha	anxious	confused	distracted	frustrated	total
Student 1	100m 38s	6	6	24	9	25	**70**
Student 2	120m 05s	4	7	15	9	4	**39**
Student 3	36m 43s	3	12	8	0	6	**29**
Student 4	37m 19s	9	18	25	4	13	**69**
Student 5	156m 02s	10	13	35	15	18	**91**
Student 6	118m 33s	2	5	23	43	5	**78**
Student 7	07m 12s	0	0	1	0	0	**1**
Student 8	58m 27s	1	2	8	8	0	**19**
Student 9	46m 16s	2	6	6	5	3	**22**
Total	**11h 21m 15s**	**37**	**69**	**145**	**93**	**74**	**418**

V/A scores were calculated from a normal distribution, given the mean score μ and standard deviation σ (Equation 3.1, 3.2). These parameters were chosen based on the steps outlined by Mollahosseini et al in their research which outlined predefined estimated regions for categorical emotions.

$$V(alence) = N[\mu + \alpha, \sigma] \quad \{ V \mid -1 <= V <= 1\} \tag{3.1}$$

$$A(rousal) = N[\mu + \beta, \sigma] \quad \{ A \mid -1 <= A <= 0.8\} \tag{3.2}$$

In AffectNet, the 'happy' emotion, for example, has a valence between (0.0, 1.0] and an arousal between [-.02, 0.5]. Annotators generally added V/A scores for 'happy' that were within that range. Jaeger et al. (2019) reported mean V/A scores from their research for several emotions including anxious, confused and frustrated. These means were normalized to achieve μ (Table 3.6). Means for aha and distracted, standard deviation for all categories were chosen based on a visual inspection of the data.

Additionally the means were adjusted by additional *a priori* weights if the tutor indicated more or less of an attribute in their annotations. One-half of a standard deviation was added to/ subtracted from the mean to indicate more or less. Occasionally, this weight was adjusted even more based on further information from the tutor.

I also postulated that given the nature and environment of this dataset (tutoring, often in a library or study hall), the tutees would likely not reach the highest threshold of arousal. Therefore maximum arousal was limited to increase variance within the research dataset. Random V/A scores were then calculated bound by the given regions and added to the dataset by a python program. Figure 3.2 shows the predefined ranges that were used in the calculations of V/A scores for observations in the tutoring dataset.

Table 3.6: V/A Mean and Standard Deviation for Ground Truth

	Valance Mean	**Valance s.d.**	**Arousal Mean**	**Arousal s.d.**
aha	.450	.250	.350	.250
anxious	-.350	.250	.220	.100
confused	-.700	.150	-.200	.100
distracted	-.200	.200	-.400	.200
frustrated	-.600	.180	.550	.180

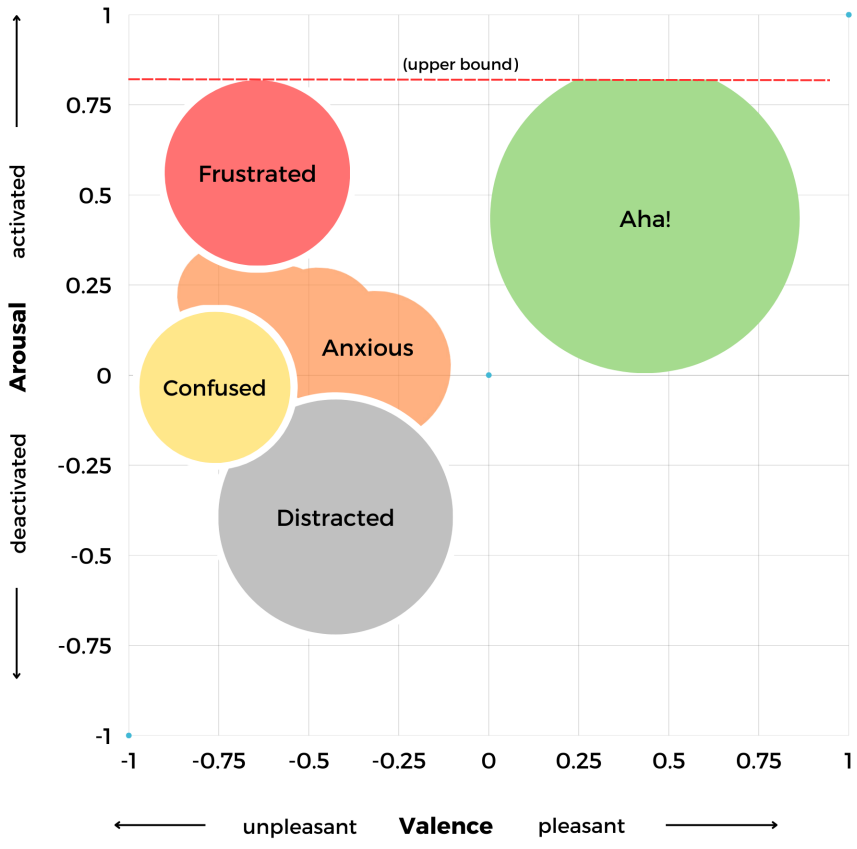


Figure 3.2: Predefined estimated valence/arousal regions.

3.8 Research Instruments

A face detection module and two FER models were the software programs used in this research. The face detection utility was used to identify faces and the FER models were used to process the recorded observations of students while they were being tutored in basic calculus.

3.8.1 RetinaFace - A Face Detection Utility

While both models had at least one facial recognition utilities embedded in their model not all were sensitive enough to capture a face if the subject's head was turned more than a few degrees. Since students often change position during tutoring (e.g., leaning down to look at a paper, or turning to look at the tutor), this study required a facial recognition utility that would recognize and capture a face at various angles. This was important for proper classification since the tutors often identified emotion classifications when tutee's face was partially turned away, and even occasionally when the face was partially obscured due to a hand scratching a face etc. RetinaFace (Serengil & Ozpinar, 2021) is a face detection program that is sensitive enough to detect faces at many angles even when they are partially obscured. Additionally, RetinaFace has a face alignment option that was used to crop faces and place them on a black background and then align faces when possible so that they are fully horizontal. RetinaFace was added to the ESR9 model as an option to use for facial recognition during the pre-processing of the data. The resulting dataset was also used in the testing of the HSEmotion model

3.8.2 ESR9 Model - A Discrete Model Classifier

The Ensemble with Shared Representations (ESR9) model chosen for this research was originally developed by researchers in the Department of Informatics at the University of

Hamburg (Siqueira et al., 2020). ESR models are so named because the convolutional neural networks (CNNs) share the information gathered between previously common layers. This particular model branches 9 times, with each new branch pursuing feature extraction slightly differently from the layers that preceded. This ESR model was designed to be trained on static images.

Image Classification

The ESR9 model required moderate changes in order to be functional for this research project. Most significant (and expected) was the need to train the model using new classifications. Current facial expression recognition research often uses a common classification system of the 6 basic emotions. This ESR model was originally developed and trained using the AffectNet dataset, which employs the common classification system. Most pre-trained discrete models available for contemporary research would use this classification system. This study was best served by using learning-centered affective states. The ESR9 model was chosen in part because the developers provided enough documentation to adequately modify the model with new discrete classifications. The model was modified using the utilities supplied by the developers with modifications to the code, and new utility programs developed for this research project (Appendix B).

Data Pre-processing

The ESR9 model was designed to be trained on the AffectNet dataset. The tutor recordings were modified so the ESR9 model could process the data as if it were AffectNet images. The bulk of this work included turning the videos into a collection of still images and then formatting the images properly. A series of Python programs were developed to prepare the recordings to be consumed by the ESR9 model. See Appendix B for detailed code.

Model Training

The ESR9 model was retrained using the dataset collected during the course of this study. During machine learning training, the algorithm uses images from the training and validation datasets to adjust weights and biases to allow the model to make the best predictions when presented with new data. The most efficient number of training epochs the developers used when training the ESR9 model on AffectNet was 50 for the first branch and 20 epochs for branches 2-9. When training on the new tutoring dataset, the number of epochs remained the same. This preserved the integrity of the loss function which is used to evaluate the model's performance.

3.8.3 HSEmotion Model - A Valence and Arousal Predictor

The High-speed emotion recognition (HSEmotion) model was the second model chosen to evaluate the recordings collected during this study (Savchenko, 2022). This model, specifically the *'enet_b0_8_va_mtl.pt'* model, took more time than most other models to carefully pre-process the training data, but resulted in an accurate model that was fully trained over 10 epochs (65% accuracy on AffectNet test data, F1 score of .609). This model output differs from ESR9 in that it uses valence and arousal for the prediction model. HSEmotion was also trained using the AffectNet dataset. Since discrete classifiers were not used for these model predictions, there was no need to retrain the model. This model was designed to accept still images or short video clips as input. However, the video processing does not produce V/A scores.

V/A Scoring

The chosen HSEmotion model offered V/A scores as part of its prediction. The model was trained using AffectNet V/A scores, which were the guidelines used to assign V/A scores to

this study's training data. This alignment ensured that the predicted V/A scores for the testing images would be comparable to the ground truth V/A scores (see Section 3.7.2).

Data Pre-processing

The images that were prepared for the ESR9 model training were used for testing the HSEmotion model as well. There were no additional changes made to the dataset except that HSEmotion required a different folder structure to read the images. All of the training, validation and testing images were combined into one folder for easy processing by the model programs.

3.9 Data Analysis

Both models developed by the original researchers included peer-reviewed methods of model development. Data analysis is primarily focused on comparing classifications or V/A scores that are predicted by the models with the ground truth annotations developed by the student tutors.

Classification Data Analysis Plan

The comparison of the categorical data was conducted using 4 standard metrics for machine learning classification models: precision, recall, F1 score and overall accuracy. The use of these metrics aided in answering research questions 1 and 3.

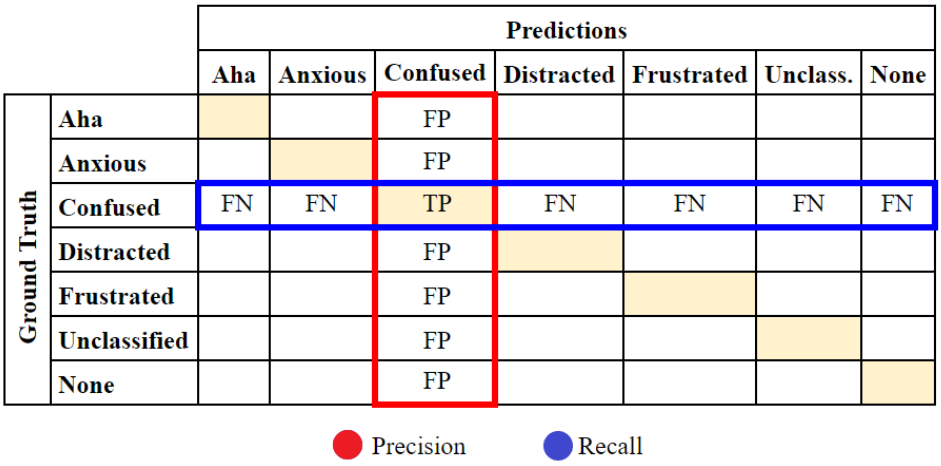


Figure 3.3: Precision and Recall.

1. **Precision** is a measure of the accuracy of the positive predictions made by the model. It is the percentage of correctly identified facial expressions in a given classification among all in that classification as identified by the model (Fig 3.3, red column). A low precision indicates that the model is making many predictions about a given category that are incorrect (false positives).

$$Precision = \frac{TruePositive}{TruePositive + TrueNegative} \tag{3.3}$$

2. **Recall** is known as the true positive rate. It is the percentage of correctly identified facial expressions in a given classification among all in that classification according to the ground truth (Fig 3.3, blue row). A low recall indicates that the model is failing to classify a significant portion of images in a given category (false negatives).

$$Recall = \frac{TruePositive}{Truepositive + FalseNegative} \tag{3.4}$$

3. **F1 score** is a calculation of predictive performance using both precision and recall. It balances the trade-offs between precision and recall to provide a single score that measures both. F1 score ranges from 0 (unable to classify any images) to 1 (every image is classified correctly).

$$F1 = \frac{Precision * Recall * 2}{Precision + Recall} \tag{3.5}$$

4. **Accuracy** is the percentage of correctly identified classifications across the entire dataset.

$$Acc = \frac{TruePositive + TrueNegative}{TotalInstances} \tag{3.6}$$

3.9.1 Valence and Arousal Score Data Analysis Plan

Measuring the accuracy of V/A scores is often viewed as more complex than discrete classifications because there is more variation in the ground truth. However, this was an important measurement for this study. The small changes in facial expression that accompany tutoring may be more identifiable using the more sensitive V/A scores. The measurement of accuracy of the predicted V/A scores was done by calculating the root-mean-square error (RMSE) of both the valence and arousal scores. RMSE is the most common metric used to measure the accuracy of continuous predictions. Reported RMSEs of other models using AffectNet ranged from .45 to .33 (Siqueira et al., 2020, p. 5807). A RMSE within that range for our training dataset indicated a comparable usage.

$$\text{RMSE} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \tag{3.7}$$

To test the null hypotheses presented in RQ2, two types of statistical tests were performed against the V/A scores as they were separated by discrete classifications. First, Pearson's correlation was calculated to test for a linear relationship between the V/A scores assigned by the tutor/researcher (the ground truth) and those calculated by the HSEmotion model. Secondly, appropriate t-tests were performed to determine if there was a significant difference in means between the ground truth V/A scores and those calculated by the HSEmotion model. This analysis was done with JMP software. Results of the analysis are shared in Chapter 4. The complete output from JMP can be found in Appendix C.

3.10 Research Validity

The measurement of the degree to which those questions were answered contributes to the validity of the study. In this study, two machine learning models were chosen that used

modern and novel approaches to deep learning. These choices were designed to give the best chance at identifying student facial expressions. Both models were trained by the creators on the largest facial expression databases available to researchers, and both performed well in testing in terms of predicting facial expressions from the classic categorizations.

The original ESR9 model had a 59% testing accuracy on the AffectNet database, which is one of the higher accuracies reported. This model was retrained during this study using the tutoring training data. While retraining the model, there was about 95% accuracy in classifying the training data and about 60% in classifying the validation data (Table 4.2). Results on the testing data are reported in Chapter 4.

In the case of the HSEmotion model, retraining was unnecessary. The creators of HSEmotion had about a 61% accuracy against predicting AffectNet classes with the *'enet_b0_8_va_mtl.pt'* model.

Model Choice

Model choice was the first step in maintaining the validity of the research. Both models chosen were peer-reviewed and used different approaches to evaluating the data. This gave the study two paths to study and assess and analyze the research questions.

In the case of this research, the study was designed to accurately measure the rate at which the two models could identify the facial expressions of the students collected in the research dataset. The training of each model was accomplished by minimizing the loss functions, as is standard for neural network processing. The purpose of a loss function is to mathematically describe how well a model evaluates a dataset. ESR9 calculated a loss function for each convolutional branch. The ensemble minimizes the combined loss functions to produce the best ensemble model. During the retraining of the ESR9 model, no changes were made to the loss function, with the exception of the hyperparameters θ to accommodate the new training image dataset. HSEmotion model used the Sharpness-Aware Minimization

(SAM) as their loss function. No changes were made to the loss function for this model. Both models achieved low loss rates and high accuracy rates after model training.

Availability of the Dataset

The availability of other researchers to use this dataset is limited. The consent form the students signed ensured that the recordings of their tutoring sessions would be destroyed at the conclusion of this study. Unfortunately this may impact the perceived validity of this study. However, it was imperative that the identity of the participants remain anonymous. A student who received tutoring may feel embarrassed or self-conscious regarding their participation in the activity. An assurance of anonymity was an important part of the recruitment process. There is, however, nothing to prevent other researchers from collecting similar data. There were no special tools used to record students, and nothing unique about the collection and storage process.

Bias in the Dataset

Another potential concern is the introduction of bias into the dataset. Steps were taken to mitigate this risk and ensure an unbiased dataset. As previously mentioned, this researcher did not stay in the vicinity where the tutoring was taking place. This was to reduce the likelihood that the student's participation in the study would impact the tutoring lesson itself. Participation bias might cause some students to react differently than they might if they were not involved in the research. While there is no way to completely eliminate that concern, steps were taken to lessen it.

Additionally, steps were taken to limit bias in the annotation of the dataset. Each tutor was allowed to thoroughly critique their own observations. They could rewind the recording, listen to the conversation, zoom in on the tutee, and take as much time as they needed to annotate each recording. Each observation noted by the annotators was independently

reviewed by myself or, in one case, another tutor. Ultimately the tutor's observations were the final word; however, that word was vetted.

3.11 Summary

The purpose of this chapter was to detail the research methods used in this study. The research design and setting was examined in detail. I gave an overview of the participants and how they were identified. I discussed the details of data collection and the deep learning models used. Finally, the method of data analysis and a discussion of the validity of the research was outlined. Chapter 4 provides the results of the analyses.

Chapter 4

Results

4.1 Introduction

This study was conducted to answer the following research questions:

RQ1. Can certain facial expressions of higher education students being tutored in basic calculus be accurately classified by a facial expression recognition deep learning model?

RQ2. Can models trained on large publicly available datasets be effectively used to assess students' emotional states during basic calculus tutoring?

To evaluate the relationship between the ground truth V/A scores and the scores computed by the model, this study began with two null hypotheses:

$H1_0$. There is no linear relationship between valence or arousal when comparing tutor defined V/A scores to the model calculated V/A scores.

$H2_0$. There is no difference in mean valence or mean arousal scores when comparing tutor defined V/A scores to the model calculated V/A scores.

RQ3. Do facial expressions alone offer enough information for convolutional neural networks to accurately assess tutee comprehension?

Video footage of students being tutored in basic calculus was collected and analyzed. The videos were reviewed by the tutors and the student's affective states were annotated, producing ground truth that was used as a baseline for all analysis in this study.

The tutoring video dataset was processed and assessed by two different machine learning algorithms. First, the ground truth was used to train an ensemble deep learning model, ESR9 (Siqueira et al., 2020). The ESR9 model attempted to classify the data into discrete learning-centered classifications. Next, the affective state observations noted by the tutors were assigned V/A scores. These scores were heavily influenced by the methodology of the score assignment of the AffectNet database (Mollahosseini et al., 2019). These scored observations were then compared to predicted V/A scores of the second model, the HSEmotion (Savchenko, 2022). Finally the discrete classifications and V/A score analysis were compared to determine if the predicted V/A scores were closely aligned with the regions on the V/A matrix that are assigned to the discrete classifications in Fig 3.2.

This chapter outlines the results of each model test and the relationship of the results to the research questions. The chapter includes details on the tutoring dataset used to train and test the models, the metrics produced, and a summary of the findings.

4.2 Summary of the Dataset

A total of 11.4 hours of tutoring video footage was collected over the course of a semester. Students were recorded in person and online. This footage was trimmed to remove moments where the student was not in the frame, or the student's face was completely obscured. This reduced the video footage to approximately 9.1 usable hours. Tutors reviewed the videos and marked the instances where they observed the tutee expressing one of the 5 learning-affective facial expressions ('aha', 'anxious', 'confused', 'distracted', 'frustrated'). The tutors identified 418 observations and the 'start' and 'end' time of the facial expressions

were marked. Total video time of annotated facial expressions was approximately 40 minutes (Table 4.1). The tutoring study dataset therefore was about 94% 'unclassified'. This was an expected outcome, since the tutors were instructed to annotate only the facial expressions indicated from the learning-centered affect list. However, that left the resulting training set very imbalanced. Subsequent steps were taken to address the imbalance during analysis.

4.3 Discrete Classification Results - ESR9 Model

The ESR9 model was designed by Siqueira et al. (2020) with the classic emotion classifications and was trained on several public facial expression databases including the AffectNet database. As a first step, the training dataset was prepared. The dataset proportions loosely followed the precedent set by the AffectNet database. AffectNet has 8 facial expression categories. The most prolific facial expression in the database is 'happy', which accounts for about 32% of the data. The least represented is 'disgust' which accounts for about 1%. In the tutoring data, the unclassified category was trimmed to become about 35% of the dataset to align with the largest category observed in AffectNet. The trimmed dataset contained approximately 127,000 images (Table 4.1, row 5).

After processing the images produced from the tutoring recordings, the testing data images (Table 4.1, row 5c) were fed through the original ESR9 model as a baseline. The purpose of this initial testing was to ensure that the tutoring image data collected would yield results if processed on the original model. While the classifications themselves would not make sense in the context of this study, processing the tutoring dataset on the original model ensured that the baseline algorithm was capable of classifying this testing data. This was an important starting point. Of the 24,983 faces captured in the 25,407 images in the training dataset, the original ESR9 model was able to classify 21,810 or 87.3% of them.

Table 4.1: Training/ Validation/ Testing Datasets

		time	**frame**
1.	Original recordings	11.4 hours	1,231,200
2.	Usable footage	9.1 hours	982,800
3.	Observations	42 mins	75,591
4.	Pre-processed image frames		982,000
5.	Original training dataset (unclassified reduced to 5%) ESR9 model only		127,033
	a. Training		81,301
	b. Validation		20,325
	c. Testing		25,407
6.	Balanced dataset (3,000 - 3,500 in each class)		23,704
	a. Training		15,409
	b. Validation		3,792
	c. Testing		4,503
7.	Extremes dataset ('aha' and 'frustrated' only)		7,360
	a. Training		3,113
	b. Validation		778
	c. Testing		975
8.	Large testing dataset (HSEmotion model only)		35,913

Next, the ESR9 model was retrained using the training and validation datasets (Table 4.1, rows 5a, 5b). The model was retrained to use the classifications outlined for this study. The model architecture remained the same as the baseline with the exception of the final classification layer, which was reduced from 10 to 7 classes (the 5 learning affect expressions along with 'unclassified' 'no face detected').

Training was done on a Dell XPS laptop with a Windows 11 operating system and Nvidia GeForce RTX 3060 Graphic Processing Unit. Training took approximately 9.5 hours to complete. The ensemble model consisted of 9 different branches. Branch 1 was trained over 50 epochs on the training images while branches 2-9 were trained with 20 epochs each. Loss and accuracy calculations are shown for the Branch 9 epochs. Generally, a machine learning algorithm uses the loss function to adjust parameters so the model can 'learn' what features are important in a training set (see Section 3.10). The progressive decrease in the loss function and continual improvement in accuracy indicate that the training on this CNN was progressing as expected (Table 4.2).

The testing data (Table 4.1, row 5c) was then fed to the model. No classifications were detected in the testing data by the retrained model. It is presumed that the prevalence of the 'unclassified' category in the training dataset led to it being the most recognized classification. Imbalanced data are a common challenge in machine learning data collection, and this imbalance can hinder the identification of minority classifications in classification models.

To address this issue, a smaller dataset was created with a more balanced distribution. Each classification contained between 3,000 and 3,500 images. This significantly reduced the amount of images in the training dataset to about 24,000 images (Table 4.1, row 6) . Training loss and accuracy results were comparable to results reported in Table 4.2. Using a smaller training dataset may introduce its own challenges. In this case, a more balanced dataset yielded slightly improved results during testing.

Table 4.2: ESR Training Loss Values

	Loss	Acc		Loss	Acc
Branch 9, Epochs 01	1.760	0.661	Branch 9, Epochs 11	0.958	0.913
Branch 9, Epochs 02	2.058	0.729	Branch 9, Epochs 12	0.874	0.914
Branch 9, Epochs 03	1.643	0.771	Branch 9, Epochs 13	0.939	0.920
Branch 9, Epochs 04	1.878	0.778	Branch 9, Epochs 14	0.902	0.919
Branch 9, Epochs 05	1.708	0.820	Branch 9, Epochs 15	1.018	0.910
Branch 9, Epochs 06	1.466	0.829	Branch 9, Epochs 16	0.834	0.935
Branch 9, Epochs 07	1.356	0.863	Branch 9, Epochs 17	0.869	0.929
Branch 9, Epochs 08	1.524	0.865	Branch 9, Epochs 18	0.896	0.936
Branch 9, Epochs 09	1.419	0.862	Branch 9, Epochs 19	0.765	0.928
Branch 9, Epochs 10	1.164	0.904	Branch 9, Epochs 20	0.651	0.951

On the second trained model, 2 of the 9 branches were able to reliably identify the 'confused' and sometimes able to identify 'anxious'. No other facial expressions were identified in any branch. Because the ensemble model used a voting system for classification, the ensemble model was unsuccessful in identifying any facial expressions. The confusion matrix and metrics are reported for Branch 9, the most accurate of the branches.

Precision measures the mistakes a model makes in its predictions while recall measures what the model missed. See Section 3.9 for more discussion of precision and recall. The precision and recall table (Table 4.3) showed a notable disparity, with a high recall indicator of 0.79 for the 'confusion' category but low precision, signaling a considerable false positive rate. This suggests frequent misidentification of other facial expressions as 'confused.' The confusion matrix (Table 4.4) corroborates this, highlighting that the model predominantly selected 'confused' in the majority of instances.

The most sensitive branch of ESR9 failed to detect 'aha' 'distracted', or 'frustrated.' Another test was proposed to evaluate how well the model would perform on identifying only these most extreme facial expressions. The model was retrained with a balanced dataset containing only expressions classified as 'aha' and 'frustrated' (Table 4.1, rows 7a, 7b).

The precision and recall rose considerably for both emotion categories (Table 4.5. But again, a single category (this time 'aha') was the dominant choice by a model that was trained on a balanced dataset. This repeated issue of bias toward a particular classification could be due to several issues:

Overfitting

The model may have overfit the data, capturing features specific to a handful of images and attempting to apply them broadly across the entire dataset. This seems improbable given that data pre-processing involved cropping images to display only the face against a black background. Visually, there appears to be little that lacks generalizability.

Table 4.3: Metrics

	Precision	Recall	F1
aha	.000	.000	.000
anxious	.201	.150	.172
confused	.183	.788	.297
distracted	.000	.000	.000
frustrated	.000	.000	.000
unclassified	.092	.020	.033
no face	.567	.982	.719
Accuracy	.250		

Table 4.4: Confusion Matrix

		ESR9 Branch 9 Predictions							
		Aha	**Anx.**	**Conf.**	**Distr.**	**Frust.**	**Unclass.**	**None**	**N's**
Ground Truth	**Aha**	.000	.207	.584	.000	.000	.106	.103	435
	Anx.	.000	.150	.674	.000	.000	.010	.166	674
	Conf.	.000	.093	.788	.000	.000	.017	.101	1158
	Distr.	.000	.074	.838	.000	.000	.052	.036	717
	Frust.	.000	.116	.653	.000	.000	.017	.215	899
	Unclas.	.000	.104	.748	.000	.000	.020	.127	393
	None	.000	.000	.019	.000	.000	.000	.982	54

Table 4.5: Metrics, Extreme Dataset

		Predicted			
		aha	frust.	Precision	.520
				Recall	.908
Ground Truth	aha	.908	.092	F1	.661
	frust.	.840	.160	Accuracy	.534

Model Issues

The model may not be well constructed or well equipped to handle the given data. Again, this seems improbable, given this was a model built specifically for classifying facial expressions. It was peer reviewed, trained on several hundred thousand diverse images, and performed well during its original testing with the authors.

Similar Classes

The data may contain many features that are highly similar to all classes, making them difficult to separate. This is likely the issue in this case. The model may default to the class it found the easiest to identify during training. This issue may be overcome by more training epochs and/or more training data. There was likely not enough data for the model to extrapolate the distinguishing features.

4.4 Valence/Arousal Scoring Results - HSEmotion Model

Valence and arousal (V/A) scoring is an alternate way to identify facial expressions in machine learning. Valence is a measure of pleasantness in a facial expression and arousal is a measure of activation. The two scores together are meant to identify the range of human emotion.

The HSEmotion model (Savchenko, 2022) was trained on several large image datasets, but was trained specifically on AffectNet to capture V/A scores. Savchenko, the original creator of the HSEmotion model, provided special python packages to download any of several pre-trained models. The *enet_b0_8_va_mtl.pt* model was used in this case. The package included python test scripts to run to ensure that the model was running correctly. Execution of the test script confirmed the model was working as expected.

The test script was modified to sequentially process a group of images from a folder and to write the results into a csv file (Appendix B). The first run of the test script processed the same testing dataset that was used to test the ESR9 model (Table 4.1, row 5c). The ground truth V/A scores were compared with the predicted scores. The results are shown in Figures 4.1 and 4.2.

For both the valence and arousal scores, the model shows small but statistically significant explanatory value when it comes to using the ground truth to predict V/A scores. The RMSE on valence was .391 and arousal was .216. These errors were comparable to the RMSE's reported by other networks predicting V/A scored using AffectNet (Lindt et al., 2019; Siqueira et al., 2020) and also comparable to the RMSE experienced by the Affectnet annotators when comparing their overall V/A scores with each other. Other studies reported that the margin of error for arousal tended to be larger than for valence (Mollahosseini et al., 2019, p. 26). Researchers presumed that while it is fairly easy to group positive or negative facial expressions, it is more difficult to assess the activation level of that emotion. In the case of this study, the predictive ability for both valence and arousal were similar, however, the arousal score was the more accurate of the two.

4.5 Comparing Observed and Predicted Valence and Arousal Scores Within Classes

To address the null hypotheses introduced in RQ2, predicted V/A scores were separated by classification. The ground truth scores were compared to the model-based predicted scores to determine if there were relationships between the two sets of scores at the categorical level. JMP software was used for all statistical tests. See Appendix C for complete JMP output.

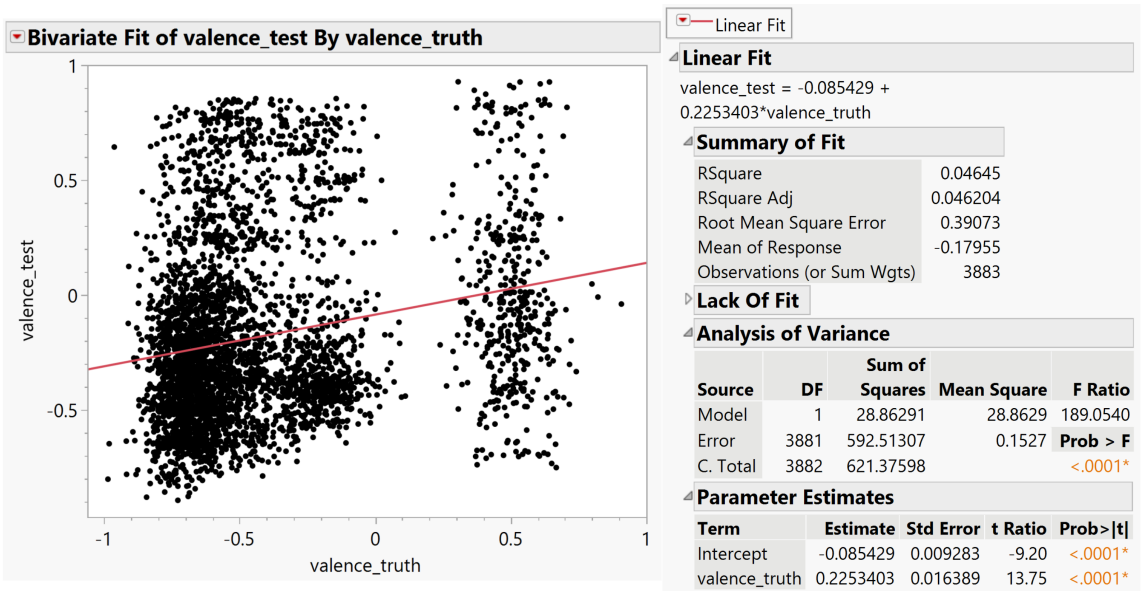


Figure 4.1: Linear Fit of Valence Observed vs Predicted

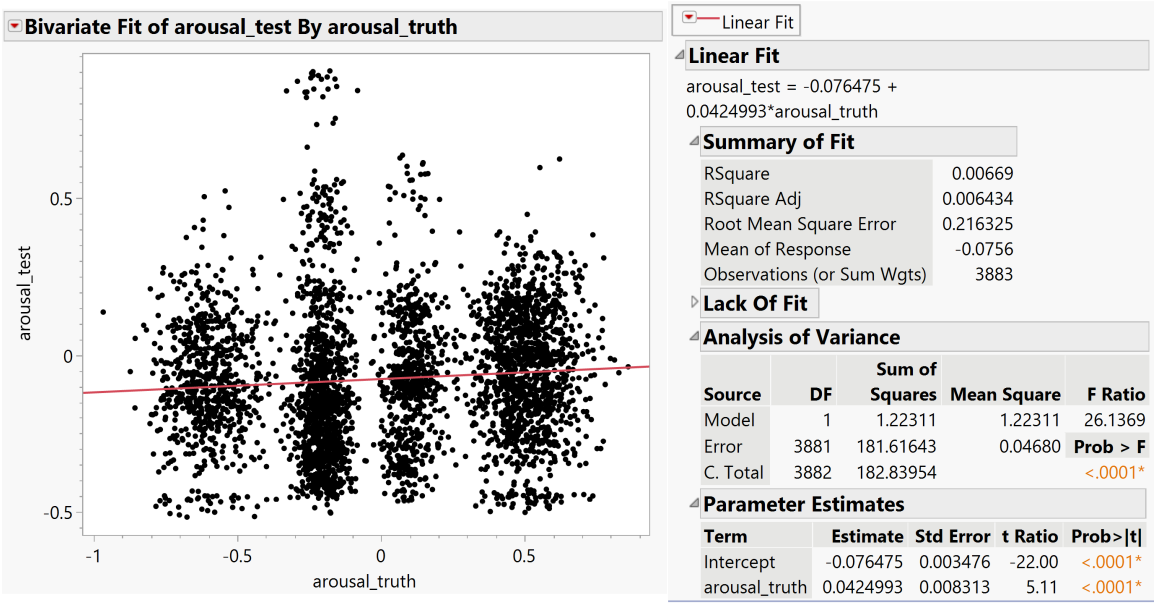


Figure 4.2: Linear Fit of Valence Observed vs Predicted

4.5.1 Linear Relationship Null Hypothesis

$H1_0$ - There is no linear relationship between valence or arousal when comparing tutor defined V/A scores to the model calculated V/A scores.

The first statistical test was used to determine if there was a linear relationship between the ground truth and model calculated scores by category. Pearson's r was used to test the possibility of correlation between the scores. Table 4.6 indicates there was no discernible linear relationship between the V/A ground truth and calculated sets of scores. The correlation coefficient for each of the classifications was nearly 0 indicating no correlation between the tutor/researcher defined ground truth and the predicted data from the model. Detailed graphs and tables are listed in Appendix C. The analysis of this data resulted in a failure to reject the null hypothesis for $H1_0$.

4.5.2 Difference in Means Null Hypothesis

$H2_0$ - There is no difference in mean valence or mean arousal scores when comparing tutor defined V/A scores to the model calculated V/A scores.

The next statistical test was used to determine if there was a relationship between the ground truth and model generated scores by category. The unequal variances indicated that Welch's t-test was the best test to use to determine mean differences. The t-test results showed that for both valence and arousal, the means of the tutor defined ground truth were significantly different from the means of the calculated scores for both valence and arousal and in every emotional category (Table 4.7). Figures 4.3 and 4.4 show a marked difference in the means and variance by way of comparative boxplots for each classification.

In the case of this study, the null hypothesis was rejected and a determination was made that for each classification, the two means were significantly different. Detailed graphs and tables are listed in Appendix C.

Table 4.6: Correlation Statistics, Ground Truth vs. Model

		Valence		Arousal	
Classification	**N's**	**Corr. Value**	**Signif. Prob.**	**Corr. Value**	**Signif. Prob.**
Aha	435	-.072	-.136	.077	.110
Anxious	674	-.064	.099	.075	.052
Confused	1158	.023	.430	-.031	.091
Distracted	717	-.064	.086	-.005	.903
Frustrated	899	.007	.830	-.027	0.427

Table 4.7: T-Test Statistics, Ground Truth vs. Model

		Valence		Arousal	
Classification	**N's**	**T-ratio**	Prob.> $\|t\|$	**T-ratio**	Prob.> $\|t\|$
Aha	435	-24.251	<.001*	-56.509	<.001*
Anxious	674	22.606	<.001*	-20.514	<.001*
Confused	1158	43.398	<.001*	12.499	<.001*
Distracted	717	4.301	<.001*	67.923	<.001*
Frustrated	899	35.347	<.001*	-76.327	<.001*

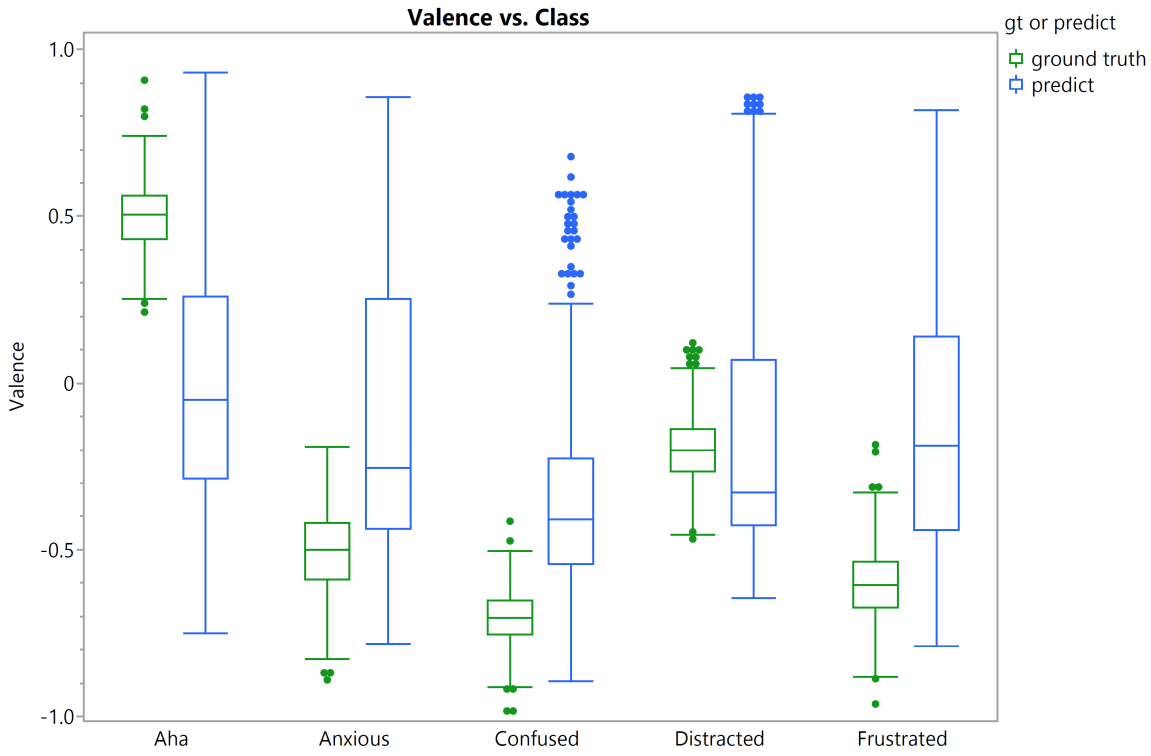


Figure 4.3: Valence Boxplot by Classification.

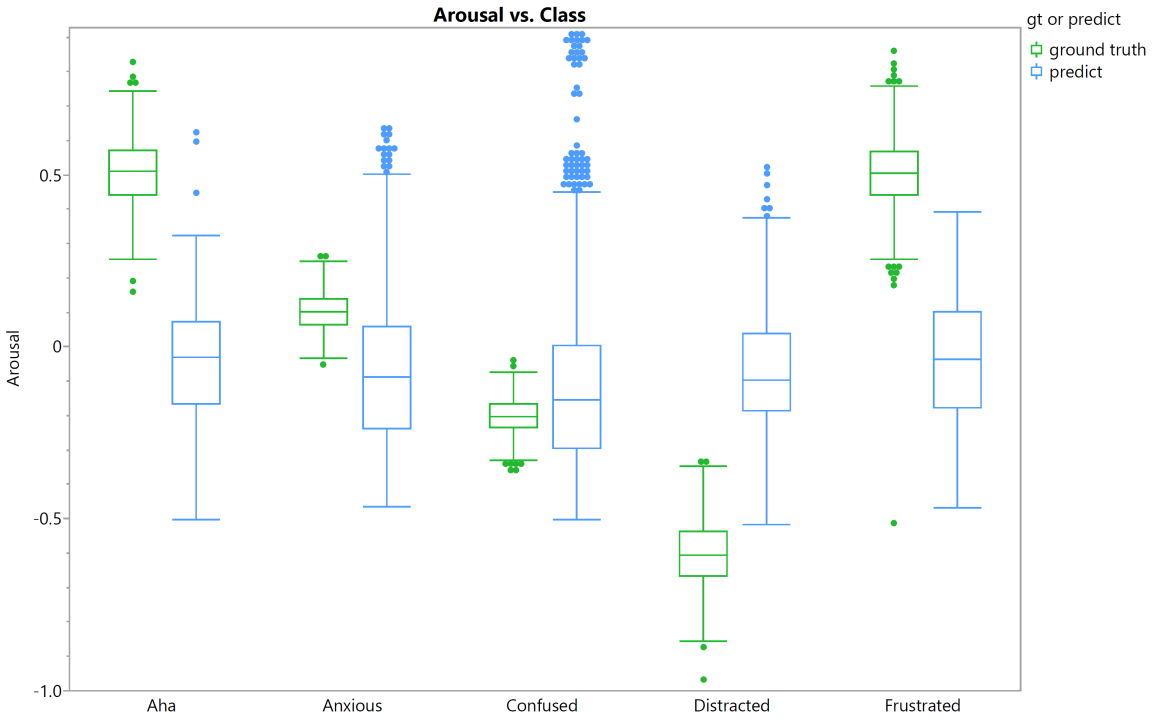


Figure 4.4: Arousal Boxplot by Classification.

4.5.3 Visual Analysis of the Data

A visual examination of the data offered more insights to the details regarding categorization. In Figure 4.5, the ground truth data is represented by the large colored circles while the predicted scores are the smaller dots. As noted in the figure, the ground truth is represented on the matrix in tight circles. The classification 'frustrated' is firmly situated in the highly unpleasant, high activation quadrant of the V/A matrix. Similarly, 'aha' is firmly established within highly pleasant, high activation. When those same images are encountered by the HSEmotion model, the predictions vary widely. 'Frustration' and 'aha' predictions both range from highly unpleasant to highly pleasant.

Because the ground truth scores were estimates, a comparison of the ground truth scores and the HSEmotion model predictions was performed by quadrant. Any prediction that had similar signs for both valence and arousal scores landed in the same quadrant as the ground truth and was considered an accurate prediction (Table 4.8, green background). Any prediction that had at least one of the two scores with the same sign landed in an adjacent quadrant and was considered mixed (Table 4.8, yellow background). Predictions that had both scores with opposite signs were considered poor (Table 4.8, red background). Following is a detailed description of the comparisons for each classification.

Aha

The observations classified as 'aha' in the ground truth represented a wide area in the V/A graph. It was the only pleasant and activated classification monitored, therefore, any observations identified as 'aha' by the tutors were assigned V/A scores between [0,0] and [1, 0.8] (Figure 4.5). V/A scores from HSEmotion accurately placed about 23% of the images in the same quadrant as the ground truth. Overall, 'aha' had the most poor predictions (Table 4.9).

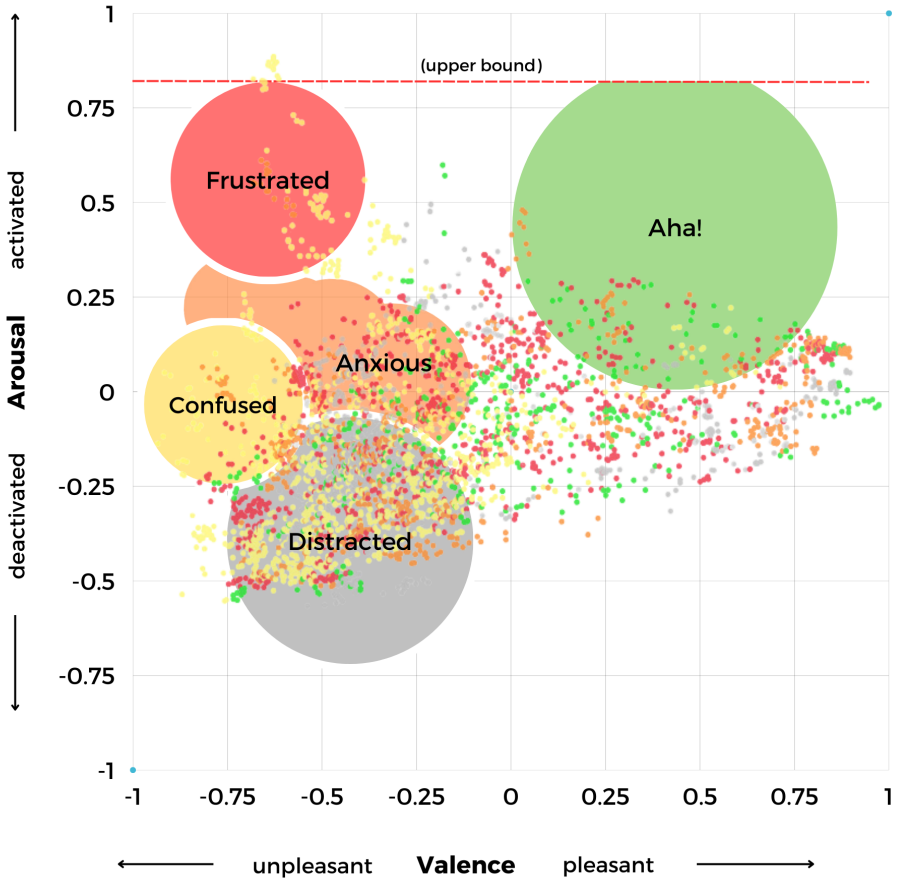


Figure 4.5: Predicted Data overlaid on Training V/A Regions

Table 4.8: Model V/A Scores by Quadrant

	q1	q2	q3	q4	Ns
Aha (+.+)	100	62	181	90	433
Anxious (-,+), (-,-)	133	90	353	98	674
Confused (-,+), (-,-)	32	267	835	24	1158
Distracted (-,-)	66	156	374	121	717
Frustrated (-,+)	172	213	379	135	899
Ns	503	788	2122	468	

Anxious

The valence and arousal scores in the tutoring dataset for 'anxious' ranged from about [-.25, .2] to [-.75, -.1] (Figure 4.5). Ground truth V/A scores for 'anxious' have mixed signs, so they span quadrants II and III. Any predictions for images labeled 'anxious' would result in either an accurate or mixed prediction. About two-thirds of the predictions were considered accurate. The remaining one third of the predictions were mixed (Table 4.9).

Confused

'Confused' V/A scores on the dataset ranged from [-.55, -.10] to about [-.80, -.30] (Figure 4.5). Like 'anxious,' the ground truth V/A scores for 'confused' span quadrants II and III. Almost all predictions for 'confused' landed in these quadrants. Confused was the facial expression most accurately predicted by both models (Table 4.9).

Distracted

'Distracted' V/A scores in the tutoring dataset ranged from about [0.0, 0.40] to [-.40, -.80]. 'Distracted' predictions landed in the ground truth quadrant about half the time. Mixed predictions were indicated about 39% of the time and poor predictions about 9% (Table 4.9).

Frustrated

The V/A scores for 'frustrated' in the tutoring dataset ranged from [-.40, .69] to [-.79, .30]. The 'frustrated' observations, like 'aha' were the most difficult for the model to predict. Only 23% were considered accurate. But unlike 'aha', the majority of the inaccurate predictions were mixed (Table 4.9).

Overall, the HSEmotion model achieved about 58% accuracy when comparing V/A score predictions to the quadrants of the ground truth.

Table 4.9: Model V/A Scores by Accuracy Classification

	Accurate Prediction	**Mixed Prediction**	**Poor Prediction**
Aha	.231	.351	.418
Anxious	.657	.343	
Confused	.952	.048	
Distracted	.522	.386	.092
Frustrated	.237	.613	.150

4.6 Summary

This chapter presented the results of this study. To answer the three research questions, two models were employed to predict facial expressions of students being tutored in basic calculus. The ESR9 model was trained first with a set of training data that contained about 35% unclassified video frames. The resulting trained model failed to identify any of the learning-centric facial expressions intended for it to learn. The model was trained a second time, with a smaller but balanced training dataset. Results were slightly better, as branch 9, the most sensitive branch of the model, identified 'confusion' and sometimes 'anxious'. Branch 8 was also able to identify 'confusion' to a lesser degree. However ESR9 is an ensemble model which employs simple voting to determine a classification of an image. The ensemble therefore was not able to identify any facial expressions.

The HSEmotion model was used to predict valence and arousal scores. Since the HSEmotion model was trained on AffectNet V/A scores, and the tutoring dataset scores were guided by the AffectNet V/A scores, there was no need to retrain the HSEmotion model. The first test was done on the same testing dataset used by the ESR9 model. Plotting a fit line on the ground truth and predicted V/A scores indicated that the model had a small but statistically significant impact when predicting V/A scores. The RSME scores of both valence and arousal were consistent with RMSE scores of other researchers testing images on an AffectNet trained model.

The model was then reviewed by classification. Again, the model was best able to identify 'confusion' based on V/A scores. Statistical analysis was also done on the V/A scores by classification to address the null hypotheses presented in RQ2. Statistical tests concluded that the observed instances of learning-affective states were significantly and observably different from the predicted V/A scores given the same data input.

Discussion and conclusions about these findings follow in Chapter 5.

Chapter 5

Conclusions

5.1 Introduction

To summarize this study, 9 participants agreed to be tutored in basic calculus on camera for research purposes. This produced over 11 hours of video footage. The video was annotated mostly by the tutors themselves, attempting to identify the moments in the tutoring session when the tutee exhibited one of 5 different learning-centric emotive responses ('aha', 'anxious', 'confused', 'distracted', 'frustrated'). The tutors annotated over 400 observations of these facial expressions. The tutoring videos were processed and still image datasets were produced. These image datasets were submitted to deep learning models as training, validation, or testing data to test whether machine learning algorithms could detect these important facial expressions as a tutor would. A deep learning discrete model, the ESR9, was retrained to classify according to the learning-centered affects. The model was retrained several times, however all attempts to classify based on the discrete classifications went poorly. A deep learning Valence/Arousal model, the HSEmotion, also was introduced which predicted valence and arousal scores. This performed moderately well on the test dataset.

This chapter will address the research questions directly and will use the results to fully answer the questions.

5.2 Summary of Findings

5.2.1 Research Question 1

RQ1. Can certain facial expressions of higher education students being tutored in basic calculus be accurately classified by a facial expression recognition deep learning model?

Answering this research question was the primary objective of this study. The facial expressions of interest were 'aha', 'anxious', 'confused', 'distracted', and 'frustrated'. Two approaches were used to address this question.

The ER9 model was used to test discrete classifications. The first training yielded no results. A second training with balanced data gave slightly improved results. The most responsive branch in the ESR9 model successfully identified 'confused' facial expressions 79% of the time. However, the recall score was low, indicating a high false positive rate. Overall, the accuracy of that branch of the model was 25%. It's also important to note that the ensemble of the branches is the intended model for ESR9. However the ensemble did not provide any results, as only 2 of the 9 branches identified any observations.

A third training was performed to assess whether the model could distinguish between the most extreme categories. This model had an increased accuracy of 54% however, bias was again observed towards one category.

Given the metrics, I determined that the ESR9 model did not accurately identify the facial expressions identified by the ground truth. There are many possible reasons for the lack of results but the most likely reason is not enough training data. The initial plan

for this study was to recruit 30 participants. Although 30 participants would have yielded better results for this model, it is unknown if even the increased number would have yielded enough training data to make this model a feasible option for use in an Intelligent Tutoring System (ITS). At the rate the available data for this study decreased (from over 1 million still images collected down to about 20,000 images used for training in the balanced dataset), it is likely that hundreds if not thousands of hours of tutoring video footage would need to be collected to effectively train a discrete classifier. In a summary answer to RQ1, the results for the discrete classifications are poor and it appears that facial expressions of students being tutored in a basic calculus course cannot be identified by this model.

5.2.2 Research Question 2

RQ2. Can models trained on large publicly available datasets be effectively used to assess students' emotional states during basic calculus tutoring?

The HSEmotion model did not report discrete classifications, but instead predicted valence and arousal (V/A) scores. The training data included V/A scores for each observation annotated by the tutors and researcher as described in Chapter 3. The V/A scores predicted by the model were compared with the V/A scores of the ground truth observations. The RMSE scores for valence and arousal (.39 and .22 respectively), while comparable to results of other studies using an AffectNet trained model, did not indicate a good prediction model for this study (Figures 4.1 and 4.2). An examination of the V/A scores when categorized by the discrete classification provided more insight into the model's utility. Given these metrics, there were indications that a deep learning FER model that focused on V/A scores could accurately identify facial expressions. However more research is required to test this theory.

It should also be noted that V/A scores for the tutoring dataset were only calculated for the moments when a facial expression was observed by the tutor. It is unknown, therefore, whether this model would identify similar V/A scores for other parts in the recording. In other words, false positives were not tested.

As indicated in Chapter 1, the facial expressions of interest span all forms of human experiences. But it is unclear if the facial expressions representing these emotions translate from one experience to another. For example, does a student who is anxious while being tutored resemble a person who is anxious while watching a movie? The method used to answer this question involves comparing the V/A scores of the observations with the predicted V/A scores calculated by the HSEmotion model.

The observed dataset represents the facial expressions that were noted and annotated by the tutors in the study. The tutors reviewed video footage and identified when they believed their tutees displayed the learning-centered affective emotions. The V/A scores assigned to the tutoring dataset were calculated based on the tutor's discrete categorizations. Regions on the V/A matrix were assigned to the discrete classes based on AffectNet V/A annotations. Expressions identified as 'aha' were assigned V/A scores within the region determined to contain 'aha' scores, and so on (Fig. 3.2). The predicted V/A scores were derived from the HSEmotion model, and represent scores that would be assigned to facial expressions observed in the wild (Kollias et al., 2019; Mollahosseini et al., 2019).

The two sets of scores were compared and the results were outlined in Chapter 4. Figure 4.5 indicates that while there was overlap when comparing the V/A scores of the observed students with the predictions of the model, the groups had no observable relationship in any classification.

The HSEmotion model had about a 57% chance of predicting V/A scores within the corresponding quadrant as compared to the ground truth. This was a broad grouping of V/A scores which served as a preliminary validation, indicating alignment with the

intended emotional context. Within this quadrant-based framework, the two V/A scores were agreement on the general emotional characterization, leaving only the intensity in question. The establishment of this broad classification suggests the HSEmotion model would work moderately well in terms of predicting V/A scores, underscoring its potential utility in capturing learning centered affects.

The most concerning problem was the high poor prediction rate for 'aha' and the low accurate prediction rates for 'aha' and 'frustrated' (Table 4.9). These were presumably the most extreme facial expressions, yet both models had trouble predicting them. This suggests that the facial expressions one might typically associate with 'frustration' or 'aha' do not always apply in the context of this tutoring dataset. A visual inspection of the data confirmed this may be the case. The following two pictures (Figure 5.1) were identified as 'frustrated' and 'aha' according to ground truth, but the predictions from the model labeled them opposite.

The overall finding of this research question is that tutoring may offer a different set of facial expression classifications than what is commonly observed in the wild (Kollias et al., 2019). When facial expressions of significance were observed during tutoring, they were often interpreted differently by a model trained on general population images.

Most of the time spent in a tutoring session will be marked by the absence of overt learning-centered facial expressions. This also is true with large facial expression databases like AffectNet which has an oversampling of 'neutral' images. The reality is we spend most of our lives wearing a 'neutral' expression (Mollahosseini et al., 2019). For an FER model to be useful in a tutoring scenario, it must be capable of detecting subtle shifts in facial expressions within this sparse environment. Adaptability to these nuanced changes will make an effective FER model in the context of tutoring sessions. In a summary to answer RQ2, The HSEmotion model which evaluated continuous scores performed moderately well. The performance of the HSEmotion model suggests that a focus on V/A scores might be

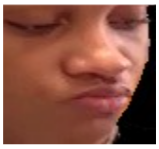
	
Ground Truth: Frustrated **Prediction**: Aha	**Ground Truth:** Aha **Prediction:** Frustrated

Figure 5.1: Extreme classifications switched

the most useful. However, there is no discernible relationship between the facial expressions of higher education students receiving tutoring in basic calculus and those observed in the broader population. In order for a FER model to be useful in a tutoring environment, a V/A model trained on tutoring images would likely be necessary.

5.2.3 Research Question 3

RQ3. Do facial expressions alone offer enough information for convolutional neural networks to accurately assess tutee comprehension?

This question represented an important secondary line of research. The machine learning models in this research study used only tutee facial expressions as input. The tutors who annotated the ground truth used additional data to inform their annotation including audio of the tutoring discussion, the tutee's body language, secondary confirmation by a researcher, and even their own memories. Given the performance of the two models, it seems feasible that with more research, facial expressions alone could be evaluated in tutoring sessions with some accuracy. However, that does not entirely address the question of whether or not facial expressions alone can provide enough information to assess tutee comprehension. This study used tutors' experiences while tutoring as the ground truth. When annotating, they were looking for facial expressions that indicated a change in tutoring tactics might be needed to help the tutee better understand the subject matter. But sometimes the tutors' experience did not match the facial expressions displayed by the tutee. The following are three examples that were observed during annotation.

1) **Student 5, recording 2, marker 00:41:49:10 (frames 75-280).** The student was nodding yes to the question "does that make sense?" During annotation, the tutor labeled the observation as 'confused'. When questioned about this, the tutor recalled "[the tutee] said she understood, but I remember that she paused for a beat before

answering so I knew she did not quite have mastery. We spent a few more minutes going over the problem until I knew she understood."

2) **Student 5, recording 3.** The tutor noted that the student was "more frustrated than usual" and showed facial expression indicators of 'frustration' several times throughout the tutoring session. However, the tutor indicated that the frustration seemed to have less to do with the tutoring session and more to do with distractions. There were other people in the tutoring center (off camera) who were working on a project and the noise level was unusually loud. While the tutor was clear that 'frustration' was the facial expression she observed, she believed distraction played a large role. The tutor was unsure if the facial expression should be labeled 'distraction' or 'frustration'. Moreover, she questioned whether the expression should be labeled at all. It was clear that the tutee was agitated, but it did not necessarily impact the tutoring session. When the tutor offered to let the student take a quick break, the tutee declined.

3) **Student 1, recording 1, marker 39:16-39:27 (frames 70,680-71,010).** The student's face was completely obscured by his hands, so those frames were excluded from the dataset. In this instance, the student presumably covered his face due to feelings of anxiety or frustration. Tutor marked this observation as anxious. The parameters of this study indicated that any frame with a totally obscured face was excluded. So those frames (and that tutor's observation), were excluded from our analysis. This is an example where body language played a much more important role than facial expressions.

These examples are information that could not be captured by facial expressions alone. In summary to answer RQ3, while facial expressions might be adequately interpreted, the issue of tutee comprehension is best assessed with as much information as is available. Suggestions

for future research regarding the nature of the additional information are included in Section 5.6.

5.3 Recommendations

Several recommendations may improve outcomes for future researchers attempting to classify facial expressions of students being tutored.

1) Control the environment as much as possible. Conducting the study in a more controlled environment could potentially lead to improved results. One unexpected and major challenge that had to be overcome during data analysis was the head angle of the student. Oftentimes, the student would turn (as if to look at a paper in front of the tutor), and the head angle would become so severe that the original face detection software could not detect a face. If the experiment was conducted in a dedicated room within a tutoring center, multiple cameras could be (unobtrusively) set up beforehand to capture a tutee's facial expressions from several different angles. If only 1 camera is available, be mindful of placement. A camera at face level is likely too high. Tutee's spend more time looking down at a book or paper or laptop than they do looking at the tutor at eye level. If the camera has good picture quality, it may be beneficial to back the camera up in order to continue capturing the face if the student shifts position.

   Also, distractions were a problem for at least one student. An environment free of interruptions and distractions would likely lead to better results. The data collection would be 'clean', free from extreme facial expressions that do not relate to the tutoring.

2) Develop clear instructions for the tutors. Though the tutors received definitions and instructions on how to classify facial expressions, there was some confusion regarding the process. There was uncertainty about whether tutors should document

every occurrence of learning-centered facial expressions or solely those instances that impacted the tutoring experience. Also, during the annotation process, tutors could denote a 'more' or 'less' attribute when they believed the tutee felt heightened or reduced levels of the emotion compared to the norm. This attribute proved crucial in assigning V/A scores. Consider revising this indicator to a confidence interval to afford tutors greater control over this attribute.

3) Organize a peer review session for annotators. In this investigation, tutors served as the primary annotators, utilizing their expertise to categorize tutees' facial expressions. Each annotator performed their annotations in collaboration with this researcher. Nevertheless, the inclusion of a peer review session would have been advantageous for tutors. Such sessions could have bolstered the confidence of some tutors in their expertise, while providing others with the valuable input of a second pair of expert eyes.

4) Focus on valence and arousal scores. The continuous V/A scores model performed moderately well despite being trained with the AffectNet general population images, which this research has shown to be an imperfect representation of the facial expressions tutees. My recommendation is that the focus should be on continuous affect scores rather than discrete classifications. V/A models like HSEmotion would likely perform even better if trained on a large tutoring images dataset.

5) Collect more video footage. The poor performance of the ESR9 model was likely an indicator of not enough training data more so than issues with the model architecture. If there is a need to conduct further research with discrete classifications, then the recommendation is to substantially increase video footage. More training epochs may have improved the performance, but more training data is key. For this study, about 94% of the video footage collected contained no observations of learning-centered

affects. If this experience is common for other researchers, then it should be understood that a substantial increase in the hours of footage collected is necessary to ensure an adequate amount of training data for the ESR9 or any other discrete model.

6) Collect more diverse video footage. Though this research project collected many hours of footage, there were only 9 faces. The initial plan was to acquire video footage from up to 30 participants, which might have limited the amount of per-person video used. But since participation was lower than expected, all recordings were used. 3 students were recorded twice and 1 student was recorded for 3 video sessions. When training a facial expression model, it is important to have a variety of facial expressions to learn from. The more diverse the video footage, the more types of facial expressions the model can learn.

7) Investigate the use of video input instead of static images. Most facial expression recognition models today use still images as input (Siqueira et al., 2020; Zhang & Ma, 2023). Those that take video input often process the video footage as independent images (Savchenko, 2022). Currently, other deep learning models are in development that process video clips as a whole and use time as a variable in emotion predictions (Deng et al., 2020; Smitha et al., 2023). This approach may be particularly useful in tutoring settings where levels of student engagement during the tutoring session may vary, causing changes to facial expressions (Lehman et al., 2008). These fluctuations could serve as an additional variable used to anticipate student needs during tutoring sessions.

5.4 Considerations

When reviewing this study, there are several considerations to take into account. First, it is difficult to classify facial expressions (Dupré et al., 2020). Everyone's own experiences and perceptions inform their interpretations of the feelings of others (Li & Deng, 2022b). This is true even in a setting as unemotional as tutoring. In this study, the annotation of the learning-centered facial expressions varied significantly among tutors. One tutor identified an emotion on average every 1-2 minutes while another tutor identified emotions every 5-6 minutes. If time permitted, having peer-review sessions where tutors reviewed each other's observations would have been useful. Keeping in mind that the ultimate goal of this research is to develop an ITS that can mimic a tutor in terms of interpreting student emotions, it is worth considering that tutors themselves vary greatly in this effort.

Another consideration is that student expressions of emotion may be different based on the tutoring experience. For example, a student who has already spent a lot of time trying to understand a problem may begin their tutoring session with the intention of getting help for that particular problem. It is feasible that their facial expressions may be notably different from another student who began a tutoring session with a more general focus. Even if they were working on the same problem, and had the same issues with understanding the concept, they may have very different feelings at the time. Both observations might be labeled as "confusion", but facial expressions of the two students might be very different. Student 2 may show significantly more arousal, as they are experiencing confusion at that moment.

Finally, there was the issue of online tutoring. This study did not distinguish between in person and online tutoring, but there are real differences and likely real changes to tutoring strategy depending upon whether tutoring is done online or in person (Curwen, 2020). Additionally, the online tutoring image quality was poor when compared to the in person videos (see differences in photo quality in Figure 5.1). While the images from online tutoring

sessions did no worse than other images of higher quality in the discrete classification, the larger impact of poor video quality in this study remains unknown. One might consider how to improve the video collected during online tutoring or consider removing it as an option.

5.5 Future Research

Several avenues for future research surfaced during the course of this study. As highlighted earlier, a key recommendation is to focus research efforts towards developing a useful V/A model. The HSEmotion model was used with pre-trained algorithms. Its performance on predicting similar V/A scores will only improve if the model is trained on tutoring-specific data. However, a focus on this model should be preceded by concentrated research on labeling V/A scores.

This study used a random classifier to assign V/A scores within the proper region of the V/A graph as outlined in Figure 3.2. The use of a more/less attribute during annotation also informed those assignments. However, a study focused solely on assigning proper V/A scores to students being tutored is an important step in order to properly train the HSEmotion model and take advantage of the continuous nature of V/A scores while focusing on the tutoring environment. These intricacies are necessary considerations to creating a powerful facial expression recognition (FER) model.

This study focused on two methods of identifying facial expressions of students during tutoring: discrete classifiers and V/A scores. Another method to consider is the Facial Action Coding System (FACS). FACS is a system used to describe facial expressions based on specific muscle movements (Rosenberg & Ekman, 2020). Extensive research has been conducted using this method since it was introduced in the 1970's. Contemporary scholars, such as Yu et al. (2021), are applying it to aid in emotion detection for ITSs. Future

research should explore whether utilizing FACS might offer comparable or superior predictive capabilities to V/A scores when classifying students' facial expressions.

Another area for future research is the incorporation of other data as input to assess tutee comprehension. The notion of relying solely on facial expressions for emotion classification, a departure from conventional practices in prior research (Mittal et al., 2020; Mukhopadhyay et al., 2020), underscores the transformative impact of deep learning on computer vision. But if the goal is to imitate a tutor, then a broad approach to collecting information is needed. Tutors use multiple senses to make quick and nuanced decisions regarding how to best meet the needs of their tutee (Hartman, 1990). While it may be prudent to focus primarily on FER, the other rich sources of information should not be discounted or forgotten.

Other questions were introduced during the course of this study that might make for interesting research in the future: 1) Does an abundance of observational data from a single student introduce bias into the training? 2) To what extent does video quality influence the training of a model to recognize learning-centered facial expressions? 3) Given that this study focused mostly on negative valence scores as a means of improving ITSs, is it worthwhile to explore the impact of positive valence expressions on tutoring? 4) Are the learning-centered facial expressions used in this research the only facial expressions tutors use when assessing student comprehension? Focused research within these avenues will enrich and advance FER research in its application to tutoring.

5.6 Conclusion

This study sought to explore whether a deep learning convolutional neural networks could imitate a tutor in classifying learning-centered facial expressions of students while they are being tutored. Two CNN models were tested. The first, the ESR9, was trained on video footage collected during this study. It attempted to classify the data into discrete

classifications. During testing, this model did not perform well. The most sensitive of the branches had a 25% accuracy. But a deeper look at precision and recall scores indicated that the accuracy score was misleading due to that branch of the model labeling most of the images as 'confused.' The ensemble model did not correctly classify any of the training data.

The second model, the HSEmotion, used V/A scores to identify facial expressions. This model was not retrained since this pre-trained model was trained on the same database that was used to identify the classification regions for the ground truth (Figure 3.2). During testing, this model performed moderately well reporting overall RMSE scores of .39 and .22 for valence and arousal (Figures 4.1 and 4.2).

The HSEmotion model also was used to compare whether the predicted V/A scores would align well with the discrete classifications identified by the tutors. Results were mixed. The category 'confusion' was fairly well predicted, but categories 'aha' and 'frustrated' both had high false positive rates. This was evidence enough to suggest that the facial expressions of students being tutored in basic calculus do not necessarily align with the facial expressions of the general population. A training of the HSE model would produce better results.

Finally, this study explored whether facial expressions alone offered enough information to assess tutee comprehension. The conclusions of this study indicate that with further research, CNN models could very well accurately identify the facial expressions of students as they are being tutored in a gateway math course. That alone, however, it is not enough to assess tutee comprehension. Tutors use a variety of input to assess student comprehension and a sole emphasis on FER risks overlooking a wealth of substantive information that may be just as accessible as facial expressions.

Although this research did not entirely attain the goal of thoroughly assessing tutee comprehension, research in the area of student facial expressions was advanced. The confirmation that facial expressions during tutoring constitute a distinct domain provides a foundation for future research on evaluating tutee comprehension. Additionally, it is

important to note that facial expressions remain the single most important predictor of student affect. On its own, precise predictions of learning-centered facial expressions would represent a significant advancement in the pursuit of effective intelligent tutoring systems.

Bibliography

Alexander, S., Sarrafzadeh, A., & Hill, S. R. (2008). Foundation of an affective tutoring system: Learning how human tutors adapt to student emotion. *International Journal of Intelligent Systems Technologies and Applications*, *4*(3/4), 355–367. https://doi.org/10.1504/IJISTA.2008.017278 (cit. on pp. 5, 25).

Alpert, D., & Bitzer, D. L. (1969). *Advances in computer-based education: A progress report on the plato program* (Report). University of Illinois. https://files.eric.ed.gov/fulltext/ED124133.pdf (cit. on p. 20).

American College Testing. (2021). *Mental health supports and academic preparedness for high school students during the pandemic* (Report). https://eric.ed.gov/?id=ED613543 (cit. on p. 19).

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *American Association for the Advancement of Science*, *228*(4698), 456–462. https://doi.org/10.1126/science.228.4698.456 (cit. on p. 1).

Arco-Tirado, J. L., Fernández-Martín, F. D., & Hervás-Torres, M. (2020). Evidence-based peer-tutoring program to improve students' performance at the university. *Studies in Higher Education*, *45*(11), 2190–2202. https://doi.org/https://doi.org/10.1080/03075079.2019.1597038 (cit. on p. 18).

Behera, A., Matthew, P., Keidel, A., Vangorp, P., Fang, H., & Canning, S. (2020). Associating facial expressions and upper-body gestures with learning tasks for enhancing intelligent tutoring systems. *International Journal of Artificial Intelligence*

in Education, *30*, 236–270. https://doi.org/10.1007/s40593-020-00195-2 (cit. on p. 24).

Benjamin, L. T. (1988). A history of teaching machines. *American Psychologist*, *43*(9), 703–712. https://doi.org/10.1037/0003-066X.43.9.703 (cit. on p. 20).

Bhatti, Y. K., Jamil, A., Nida, N., Yousaf, M. H., Viriri, S., & Velastin, S. A. (2021). Facial expression recognition of instructor using deep features and extreme learning machine. *Computational Intelligence and Neuroscience*, *2021*(5570870), 17. https://doi.org/10.1155/2021/5570870 (cit. on p. 24).

Bloom, B. S. (1984). The 2 sigma problem: The searvh for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004 (cit. on p. 19).

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 379–388. https://doi.org/10.1145/2678025.2701397 (cit. on p. 30).

Channing, J., & Okada, N. C. (2020). Supplemental instruction and embedded tutoring program assessment: Problems and opportunities. *Community College Journal of Research and Practice*, *44*(4), 241–247. https://doi.org/10.1080/10668926.2019.1575777 (cit. on p. 18).

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237–248. https://doi.org/10.3102/00028312019002237 (cit. on p. 17).

Corbishley, J. B., & Truxaw, M. P. (2010). Mathematical readiness of entering college freshmen: An exploration of perceptions of mathematics faculty: Mathematical readiness of freshmen. *School science and mathematics*, *110*(2), 71–85. https://doi.org/10.1111/j.1949-8594.2009.00011.x (cit. on p. 19).

Curwen, M. S. (2020). Vexations and breakthroughs: Taking an in-person tutoring program online. *Issues in teacher education*, *29*(1-2), 75–94 (cit. on p. 92).

Dang, H.-A., & Rogers, F. H. (2008). The growing phenomenon of private tutoring: Does it deepen human capital, widen inequalities, or waste resources? *The World Bank Research Observer*, *23*, 161–200. https://www.jstor.org/stable/40282371 (cit. on pp. 15, 16).

Darwin, C., & Prodger, P. (1998). *The expression of emotion in man and animals* (3rd). Oxford University Press. https://pure.mpg.de/rest/items/item_2309885/component/file_2309884/content (cit. on pp. 13, 23).

deBarros, A., & Ganimian, A. J. (2023). Which students benefit from computer-based individualized instruction? experimental evidence from public schools in india. *Journal of Research on Educational Effectiveness*. https://doi.org/10.1080/19345747.2023.2191604 (cit. on p. 21).

Deng, D., Chen, Z., Zhou, Y., & Shi, B. (2020). Mimamo net: Integrating micro- and macro-motion for video emotion recognition. *AAAI Conference on Artificial Intelligence*, *34*, 2621–2628. https://doi.org/10.1609/aaai.v34i03.5646 (cit. on pp. 91, 112).

deRee, J., Maggioni, M. A., Paulle, B., Rossignoli, D., Ruijs, N., & Walentek, D. (2023). Closing the income-achievement gap? experimental evidence from high-dosage tutoring in dutch primary education. *Economics of Education Review*, *94*(102383). https://doi.org/10.1016/j.econedurev.2023.102383 (cit. on p. 19).

Devi, B., & Preetha, M. M. S. J. (2021). A descriptive survey on face (facial) emotion recognition techniques. *International Journal of Image and Graphics*, *23*(1), 25. https://doi.org/10.1142/S0219467823500080 (cit. on p. 24).

D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., El Kaliouby, R., Picard, R., et al. (2008). Autotutor detects and responds to learners

affective and cognitive states. *Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems*, 306–308 (cit. on p. 26).

D'Mello, S. K., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001 (cit. on p. 24).

Dupré, D., Krumhuber, E. G., Küster, D., McKeown, G. J., Contributor, & D'Mello, S. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PloS one*, *15*(4), p.e0231968–e0231968. https://doi.org/10.1371/journal.pone.0231968 (cit. on pp. 27, 92).

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, *19*, 207–283 (cit. on p. 13).

Ekman, P., & Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and Emotion*, *10*(2), 159–168. https://link.springer.com/content/pdf/10.1007/BF00992253.pdf (cit. on p. 23).

Feng, S., Magana, A., & Kao, D. (2021). A systematic review of literature on the effectiveness of intelligent tutoring systems in stem. *2021 IEEE Frontiers in Education Conference (FIE)*. https://doi.org/10.1109/FIE49875.2021.9637240 (cit. on p. 22).

Foshay, R. (2004). *An overview of the research base of plato* (Report). PLATO Learning Incorporated. https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/PLATO_US/P040800F.pdf (cit. on p. 20).

Fryer, R. G., Jr. (2016). *The production of human capital in developed countries: Evidence from 196 randomized field experiments* (Report). National Bureau of Economic Research. http://www.nber.org/papers/w22130 (cit. on p. 19).

Gamlath, S. (2022). Peer learning and the undergraduate journey: A framework for student success. *Higher education research and development*, *41*(3), 699–713. https://doi.org/DOI:10.1080/07294360.2021.1877625 (cit. on p. 17).

Hao, K. (2019). China has started a grand experiment in ai education. it could reshape how the world learns. *MIT Technology Review*, *123*(1). https://fully-human.org/wp-content/uploads/2019/09/China-AI-experiment-education.pdf (cit. on p. 22).

Hartman, H. J. (1990). Factors affecting the tutoring process. *Journal of Developmental Education*, *14*(2), 2–4. https://www.jstor.org/stable/42775507 (cit. on pp. 16, 17, 94).

Ito, H., Kasai, K., Nishiuchi, H., & Nakamuro, M. (2021). Does computer-aided instruction improve children's cognitive and noncognitive skills? *Asian Development Review*, *38*(1), 98–118. https://doi.org/10.1162/adev_a_00159 (cit. on p. 21).

Jaafar, R., Toce, A., & Polnariev, B. A. (2016). A multidimensional approach to overcoming challenges in leading community college math tutoring success. *Community College Journal of Research and Practice*, *40*(6). https://doi.org/10.1080/10668926.2015.1021406 (cit. on pp. 19, 32).

Jaeger, S. R., Roigard, C. M., Jin, D., Vidal, L., & Ares, G. (2019). Valence, arousal and sentiment meanings of 33 facial emoji: Insights for the use of emoji in consumer research. *Food research international*, *119*, 895–907. https://doi.org/10.1016/j.foodres.2018.10.074 (cit. on p. 43).

Janice, A., & Voight, M. (2016). *Toward convergence: A technical guide for the postsecondary metrics framework* (Report). Institute for Higher Education Policy. https://www.ihep.org/publication/toward-convergence-a-technical-guide-for-the-postsecondary-metrics-framework/ (cit. on p. 31).

Jones, E. A., Voorhees, R. A., & Paulson, K. (2002). *Defining and assessing learning: Exploring competency-based initiatives* (Government Document). U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubs2002/2002159.pdf (cit. on p. 16).

KC, K., Mbanaso, A., Chosie, K., & Segui, E. (2020). Efficiency of peer tutoring program at the alabama college of osteopathic medicine (acom) [version 1]. *MedEdPublish*, *9*, 157. https://doi.org/10.15694/mep.2020.000157.1 (cit. on pp. 17, 18).

Klein, J., Moon, Y., & Picard, R. W. (1999). This computer responds to user frustration. *CHI EA '99*, 242–243 (cit. on p. 11).

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., & Zafeiriou, S. (2019). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International journal of computer vision*, *127*(6-7), 907–929. https://doi.org/10.1007/s11263-019-01158-4 (cit. on pp. 26, 84, 85).

Kosovich, J. J., Hulleman, C. S., Phelps, J., & Lee, M. (2019). Improving algebra success with a utility-value intervention. *Journal of Developmental Education*, *42*(2), 2–10. http://www.jstor.org/stable/45221757 (cit. on p. 12).

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of educational research*, *86*(1). https://doi.org/10.3102/0034654315581420 (cit. on p. 2).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/https://doi.org/10.1038/nature14539 (cit. on p. 2).

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What are you feeling? investigating student affective states during expert human tutoring sessions. *International Conference on Intelligent Tutoring Systems*, *5091*, 50–59. https://doi.org/10.1007/978-3-540-69132-7_10 (cit. on pp. 5, 24, 25, 30, 37, 91).

Li, S., & Deng, W. (2022a). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, *13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446 (cit. on p. 30).

Li, S., & Deng, W. (2022b). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, *13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446 (cit. on p. 92).

Lindt, A., Barros, P., Siqueira, H., & Wermter, S. (2019). Facial expression editing with continuous emotion labels. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–8 (cit. on p. 67).

Lisetti, C. L., & Schiano, D. J. (2000). Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics & Cognition*, *8*(1), 185–235 (cit. on p. 26).

Maxwell, M. (1990). Does tutoring help? a look at the literature. *Review of Research in Developmental Education*, *7*(4) (cit. on p. 16).

Miller, L. K. (2020). Can we change their minds? investigating an embedded tutor's influence on students' mindsets and writing. *The Writing Center Journal*, *28*(1/2), 103–130. https://www.jstor.org/stable/10.2307/27031265 (cit. on p. 18).

Milne, S., Cook, J., Shiu, E., & McFadyen, A. (1997). Adapting to learner attributes: Experiments using an adaptive tutoring system. *Educational Psychology*, *17*(1-2), 141–155. https://doi.org/10.1080/0144341970170110 (cit. on p. 22).

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Thirty-Fourth AAAI Conference on Artificial Intelligence*, *34*, 1359–1367. https://doi.org/10.1609/aaai.v34i02.5492 (cit. on p. 94).

Modén, M., Marie Utterberg Tallvid, Lundin, J., & Lindström, B. (2021). Intelligent tutoring systems: Why teachers abandoned a technology aimed at automating teaching processes. *54th Hawaii International Conference on System Sciences*. http://hdl.handle.net/10125/70798 (cit. on p. 23).

Mollahosseini, A., Behzad, H., & Mahoor, M. H. (2019). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. %7Bhttps://ieeexplore.ieee.org/document/8013713%7D (cit. on pp. 9, 27, 39, 41, 57, 67, 84, 85, 112).

Moore, R., & Hayes, S. (2023). *State of student learning in 2023* (Report). Curriculum Associates Research. https://www.curriculumassociates.com/research-and-efficacy/annual-report-the-state-of-student-learning-in-2023 (cit. on p. 19).

Morales-Rodríguez, M. L., Ramírez-Saldivar, J. A., Hernández-Ramírez, A., Sánchez-Solís, J. P., & Martínez-Flores, J. A. (2012). Architecture for an intelligent tutoring system that considers learning styles. *Research in Computing Science*, *47*, 37–47 (cit. on p. 1).

Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P. K. D., Dasgupta, N., & Choudhury, P. (2020). Facial emotion detection to assess learner's state of mind in an online learning system. *5th International Conference on Intelligent Information Technology*. https://doi.org/10.1145/3385209.3385231 (cit. on p. 94).

Nihan Er, S. (2018). Mathematics readiness of first-year college students and missing necessary skills: Perspectives of mathematics faculty. *Journal of Further and Higher Education*, *21*(1), 937–952. https://doi.org/10.1080/0309877X.2017.1332354 (cit. on p. 19).

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, *4*, 251–277. https://doi.org/10.1007/BF00168958 (cit. on p. 22).

Orey, M. (1993). Three years of intelligent tutoring evaluation: A summary of findings. *American Educational Research Association*, 1–12. https://eric.ed.gov/?id=ED357065 (cit. on p. 22).

Pai, K.-C., Kuo, B.-C., Liao, C.-H., & Liu, Y.-M. (2021). An application of chinese dialogue-based intelligent tutoring system in remedial instruction for mathematics learning.

Educational Psychology, *41*(2), 1–16. https://doi.org/10.1080/01443410.2020.1731427 (cit. on p. 22).

Person, N. K., & Graesser, A. C. (2003). Fourteen facts about human tutoring: Food for thought for its developers. *AIED Workshop on Tutorial Dialogue*, 335–344 (cit. on p. 25).

Picard, R. W. (2000). *Affective computing* (2nd ed.). MIT Press. (Cit. on p. 24).

Picard, R. W. (2010). Affective computing: From laughter to ieee. *IEEE transactions on affective computing*, *1*(1), 11–17. https://doi.org/10.1109/T-AFFC.2010.10 (cit. on p. 3).

Posner, J., Russell, J. A., & Petersona, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340 (cit. on p. 39).

Rosenberg, E., & Ekman, P. (2020). *What the face reveals: Basic and applied studes of spontaneous expression using the facial action coding system (facs)*. Oxford University Press. (Cit. on pp. 27, 93).

Russel, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714 (cit. on p. 39).

Savchenko, A. V. (2022). Hsemotion: High-speed emotion recognition library. *Software Impacts*, *14*(100433). https://doi.org/https://doi.org/10.1016/j.simpa.2022.100433 (cit. on pp. 6, 48, 57, 66, 91, 112).

Seo, E. H., & Kim, M. J. (2019). The effect of peer tutoring for college students: Who benefits more from peer tutoring, tutors or tutees? *The New Educational Review*, *58*, 97–106. https://doi.org/10.15804/tner.19.58.4.07 (cit. on p. 17).

Serengil, S. I., & Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697 (cit. on p. 46).

Siqueira, H., Magg, S., & Wermter, S. (2020). Efficient facial feature learning with wide ensemble-based convolutional neural networks. *AAAI Conference on Artificial Intelligence*, *34*, 5800–5809. https://doi.org/10.1609/aaai.v34i04.6037 (cit. on pp. 5, 47, 52, 57, 58, 67, 91, 112).

Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. Academic Press. https://shs.hal.science/hal-00702997/ (cit. on p. 21).

Smitha, E. S., Sendhilkumar, S., & Mahalakshmi, G. S. (2023). Ensemble convolution neural network for robust video emotion recognition using deep semantics. *Scientific programming*, *2023*, 1–21. https://doi.org/10.1155/2023/6859284 (cit. on p. 91).

Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S. H., AlGhamdi, M. A., Nagra, A. A., & Masood, K. (2022). Racial identity-aware facial expression recognition using deep convolutional neural networks. *Applied Sciences*, *12*(1), 88. https://doi.org/10.3390/app12010088 (cit. on p. 13).

Sparks, S. D. (2023). The state of school tutoring, in charts. https://www.edweek.org/leadership/the-state-of-school-tutoring-in-charts/2023/02 (cit. on p. 16).

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331–347. https://doi.org/10.1037/a0034752 (cit. on pp. 3, 23).

Ticknor, C. S., Shaw, K. A., & Howard, T. (2014). Assessing the impact of tutorial services. *Journal of College Reading and Learning*, *45*(1), 52–66. https://doi.org/10.1080/10790195.2014.949552 (cit. on p. 17).

Topping, J. K. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Kluwer Academic Publishers*, *32*(3), 321–345. https://www.jstor.org/stable/3448075 (cit. on p. 17).

Topping, K. J. (2013). Trends in peer learning. *Developments in Educational Psychology*, 53–68. https://doi.org/10.1080/01443410500345172 (cit. on pp. 16, 17).

Towns, S. G., FitzGerald, P. J., & Lester, J. C. (1998). Visual emotive communication in lifelike pedagogical agents. *Intelligent Tutoring Systems*, 474–483. https://doi.org/10.1007/3-540-68716-5_53 (cit. on p. 26).

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369 (cit. on pp. 22, 23).

VanLehn, K., Burleson, W., Girard, S., Chavez-Echeagaray, M. E., Gonzalez-Sanchez, J., Hidalgo-Pontet, Y., & Zhang, L. (2014). The affective meta-tutoring project: Lessons learned. *International Conference on Intelligent Tutoring Systems*, *8478*, 84–93. https://doi.org/10.1007/978-3-319-07221-0_11 (cit. on p. 22).

Woolley, D. R. (2016). Plato: The emergence of online community. In *Social media archelogy and poetics*. MIT Press. https://lccn.loc.gov/2015042571 (cit. on p. 20).

Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on k-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, *50*(6), 3119–3137. https://doi.org/http://dx.doi.org/10.1111/bjet.12758 (cit. on p. 3).

Yu, H., Gupta, A., Lee, W., Arroyo, I., Betke, M., Allesio, D., Murray, T., Magee, J., & Woolf, B. P. (2021). Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems. *International Conference on Human-Computer Interaction*, 219–233. https://doi.org/10.1007/978-3-030-77873-6_16 (cit. on pp. 24, 27, 93).

Zhang, R., & Ma, R. (2023). Facial expression recognition method based on psa-yolo network. *Frontiers in neurorobotics*, *16*, 1057983–1057983. https://doi.org/10.3389/fnbot.2022.1057983 (cit. on p. 91).

Appendices

A IRB Approval Letter

THE UNIVERSITY OF TENNESSEE KNOXVILLE

November 09, 2022

Kriss Gordon Gabourel,
UTK - College of Engineering - Bredesen Center for Interdisciplinary Research and Graduate Education

Re: UTK IRB-21-06557-XP
Study Title: Classifying Facial Expressions of Students Being Tutored in a Gateway College Math Course

Dear Kriss Gordon Gabourel:

The UTK Institutional Review Board (IRB) reviewed your application for **revision** of your previously approved project, referenced above. It determined that your application is eligible for **expedited** review under 45 CFR 46.110(b). In addition, the Board determined that approval of your revision application is dependent on a satisfactory response to the following **administrative stipulations**.

You must respond to the following provisos using the PI Response to Review form found in your "All Tasks" and labeled as a "Submission Response" located in the iMedRIS system online. Please use the PI Response to Review form to create any necessary revisions to study documents. Call the IRB at (865) 974-7697 with any questions.

Submission stipulations

1. Please create a revised version of the IRB application and make these changes to participant compensation there as well.

Further review by the IRB is contingent upon submission of a satisfactory response.

Sincerely,

Lora Beebe, Ph.D., PMHNP-BC, FAAN
Chair

Institutional Review Board | Office of Research & Engagement
1534 White Avenue Knoxville, TN 37996-1529
865-974-7697 865-974-7400 fax irb.utk.edu

B Repositories and Code Packages

B.1 Dissertation Research Code

Classifying Facial Expressions of Students Being Tutored in a Gateway College Math Course

Github: https://github.com/krissgab?tab=projects

B.2 ESR9 Model Code

Efficient Facial Feature Learning with Wide Ensemble based Convolutional Neural Network (Siqueira et al., 2020)

Github: https://github.com/siqueira-hc/Efficient-Facial-Feature-Learning-with-Wide-Ensemble-based-Convolutional-Neural-Networks

B.3 HSEmotion Model Code

HSEmotion Python Library for Facial Emotion Recognition (Savchenko, 2022)

Github: https://github.com/HSE-asavchenko/hsemotion

B.4 RetinaFace

RetinaFace" RetinaFace: Single-shot Multi-level Face Localisation in the Wild (Deng et al., 2020)

Github: https://github.com/deepinsight/insightface

B.5 Affectnet

AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild (Mollahosseini et al., 2019)

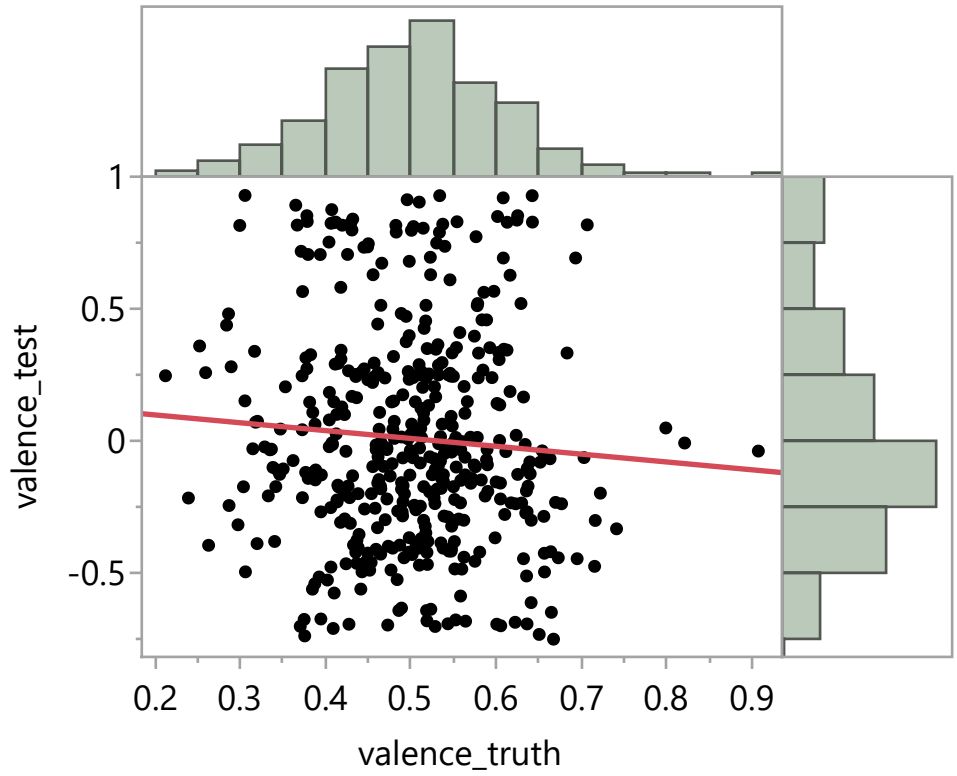
URL: http://mohammadmahoor.com/affectnet

C JMP Statistical Output

C.1 Correlation Analysis of Observed and Predicted Valence Scores within Classes

The following tables and graphs represent the analysis of linear relationships between ground truth and model calculated valence scores when the observations are grouped by tutor defined classifications. Note that the x-axis, labeled as 'valence_truth' marks observed scores, or ground truth as identified by the tutors. The y-axis, labeled 'valence_test', are the valence scores predicted by the model.

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Aha



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.07161	-0.16452	0.022565	0.1359
Covariance	-0.00292			
Count	435			

Variable	Mean	Std Dev
valence_truth	0.500258	0.099137
valence_test	0.008421	0.411207

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Aha

Linear Fit

valence_test = 0.1570052 - 0.2970154*valence_truth

Summary of Fit

RSquare	0.005128
RSquare Adj	0.00283
Root Mean Square Error	0.410624
Mean of Response	0.008421
Observations (or Sum Wgts)	435

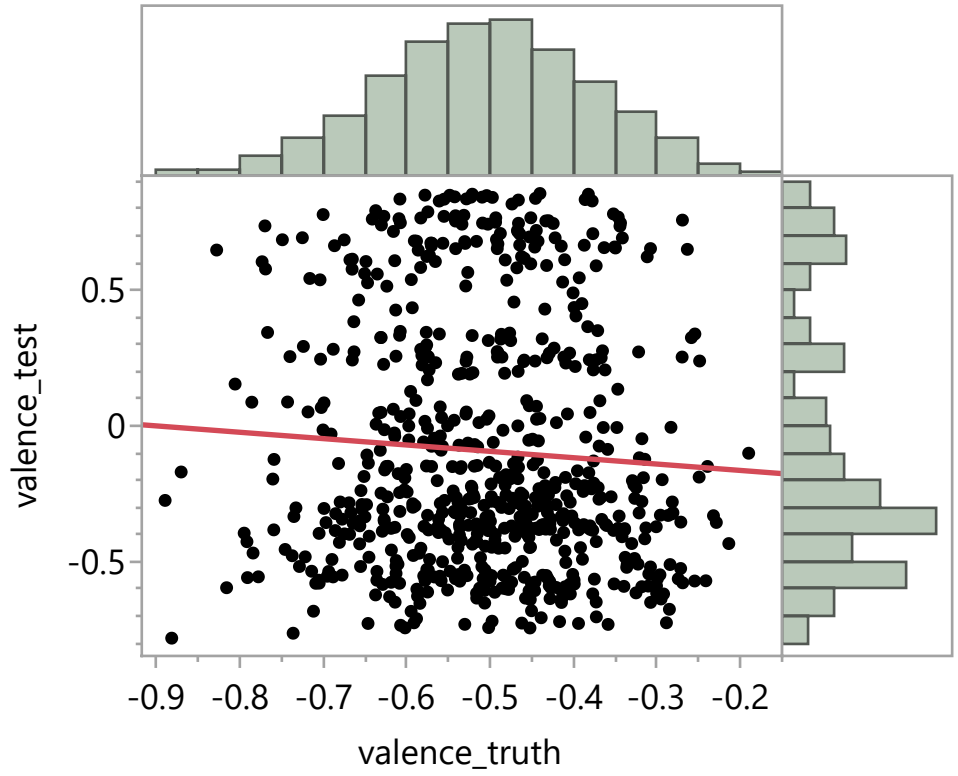
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.376288	0.376288	2.2317
Error	433	73.009182	0.168612	**Prob > F**
C. Total	434	73.385469		0.1359

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.1570052	0.101392	1.55	0.1222
valence_truth	-0.297015	0.198822	-1.49	0.1359

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Anxious



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.06359	-0.13844	0.011989	0.0991
Covariance	-0.00362			
Count	674			

Variable	Mean	Std Dev
valence_truth	-0.50552	0.124265
valence_test	-0.09262	0.457612

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Anxious

Linear Fit

valence_test = -0.211 - 0.2341673*valence_truth

Summary of Fit

RSquare	0.004043
RSquare Adj	0.002561
Root Mean Square Error	0.457026
Mean of Response	-0.09262
Observations (or Sum Wgts)	674

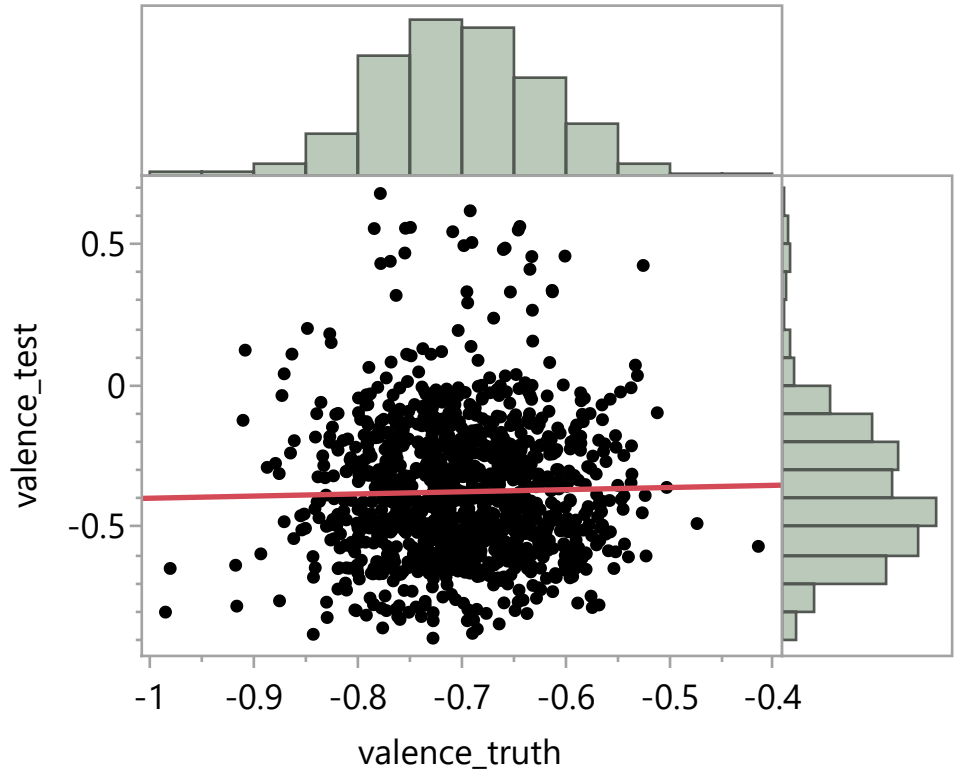
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.56985	0.569853	2.7282
Error	672	140.36234	0.208873	**Prob > F**
C. Total	673	140.93219		0.0991

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.211	0.073799	-2.86	0.0044*
valence_truth	-0.234167	0.14177	-1.65	0.0991

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Confused



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.023206	-0.03445	0.080705	0.4302
Covariance	0.000422			
Count	1158			

Variable	Mean	Std Dev
valence_truth	-0.70198	0.07454
valence_test	-0.37684	0.243811

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Confused

Linear Fit

valence_test = -0.323558 + 0.0759023*valence_truth

Summary of Fit

RSquare	0.000538
RSquare Adj	-0.00033
Root Mean Square Error	0.243851
Mean of Response	-0.37684
Observations (or Sum Wgts)	1158

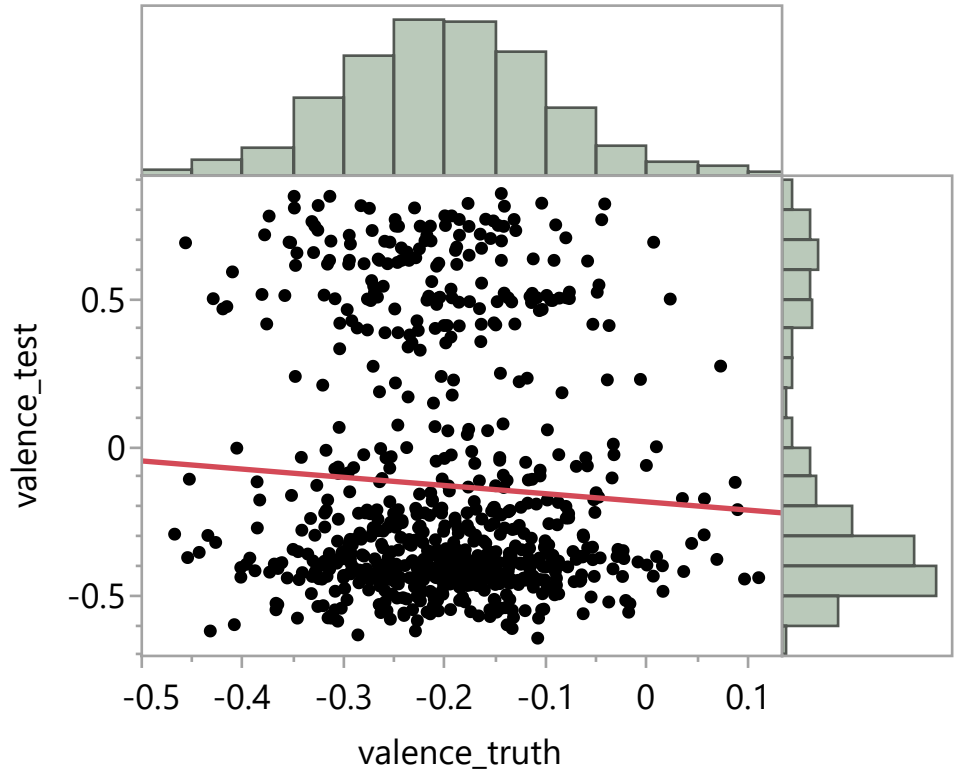
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.037036	0.037036	0.6228
Error	1156	68.739601	0.059463	**Prob > F**
C. Total	1157	68.776637		0.4302

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.323558	0.067893	-4.77	<.0001*
valence_truth	0.0759023	0.096176	0.79	0.4302

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Distracted



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.0641	-0.13667	0.009166	0.0863
Covariance	-0.00264			
Count	717			

Variable	Mean	Std Dev
valence_truth	-0.19912	0.097709
valence_test	-0.12967	0.421197

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Distracted

Linear Fit

valence_test = -0.18468 - 0.2762959*valence_truth

Summary of Fit

RSquare	0.004108
RSquare Adj	0.002715
Root Mean Square Error	0.420625
Mean of Response	-0.12967
Observations (or Sum Wgts)	717

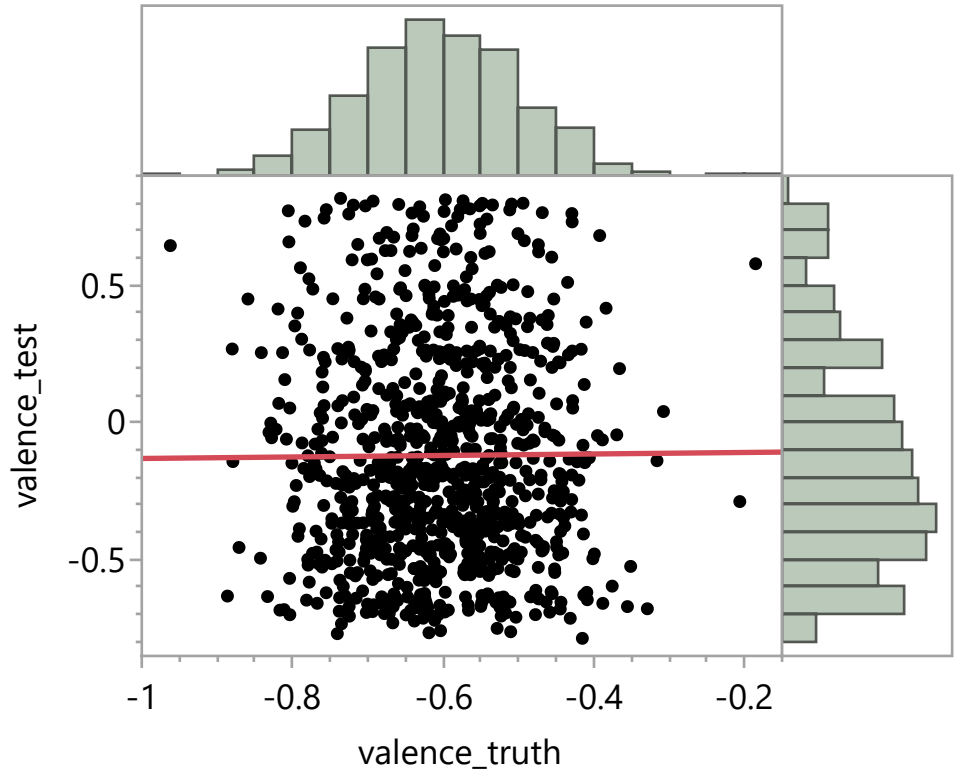
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.52183	0.521835	2.9495
Error	715	126.50143	0.176925	**Prob > F**
C. Total	716	127.02327		0.0863

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.18468	0.035678	-5.18	<.0001*
valence_truth	-0.276296	0.16088	-1.72	0.0863

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Frustrated



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.00717	-0.05824	0.072521	0.8300
Covariance	0.000295			
Count	899			

Variable	Mean	Std Dev
valence_truth	-0.60404	0.10384
valence_test	-0.12131	0.396096

Bivariate Fit of valence_test By valence_truth class_truth_lbl=Frustrated

Linear Fit

valence_test = -0.104792 + 0.0273518*valence_truth

Summary of Fit

RSquare	5.142e-5
RSquare Adj	-0.00106
Root Mean Square Error	0.396307
Mean of Response	-0.12131
Observations (or Sum Wgts)	899

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.00724	0.007244	0.0461
Error	897	140.88211	0.157059	**Prob > F**
C. Total	898	140.88935		0.8300

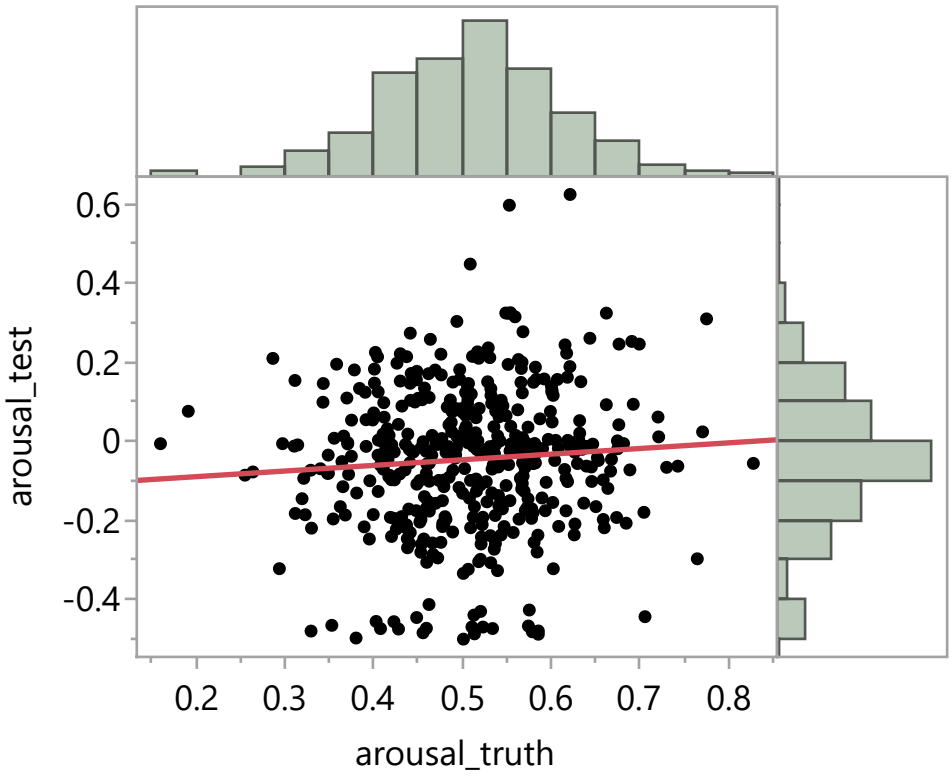
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.104792	0.078057	-1.34	0.1798
valence_truth	0.0273518	0.127359	0.21	0.8300

C.2 Correlation Analysis of Observed and Predicted Arousal Scores within Classes

The following tables and graphs represent the analysis of linear relationships between ground truth and model calculated arousal scores when the observations are grouped by classifications. Note that the x-axis, labeled 'arousal_truth' marks the observed scores, or ground truth as identified by the tutors. The y-axis, labeled 'arousal_test', are the arousal scores predicted by the model.

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Aha



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.076664	-0.01748	0.169463	0.1103
Covariance	0.00134			
Count	435			

Variable	Mean	Std Dev
arousal_truth	0.508569	0.09694
arousal_test	-0.04608	0.180305

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Aha

Linear Fit

arousal_test = -0.118595 +
0.1425933*arousal_truth

Summary of Fit

RSquare	0.005877
RSquare Adj	0.003582
Root Mean Square Error	0.179982
Mean of Response	-0.04608
Observations (or Sum Wgts)	435

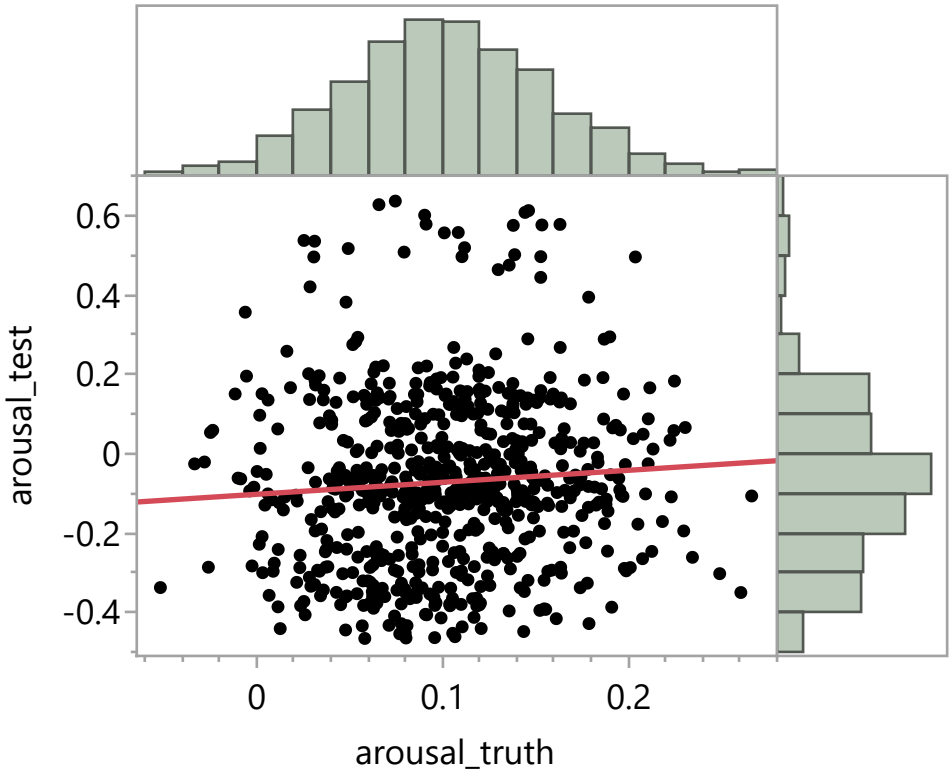
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.082926	0.082926	2.5600
Error	433	14.026344	0.032393	**Prob > F**
C. Total	434	14.109270		0.1103

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.118595	0.046139	-2.57	0.0105*
arousal_truth	0.1425933	0.089121	1.60	0.1103

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Anxious



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.074837	-0.00069	0.149512	0.0521
Covariance	0.000843			
Count	674			

Variable	Mean	Std Dev
arousal_truth	0.101997	0.052831
arousal_test	-0.07166	0.213323

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Anxious

Linear Fit

arousal_test = -0.102481 + 0.3021813*arousal_truth

Summary of Fit

RSquare	0.005601
RSquare Adj	0.004121
Root Mean Square Error	0.212883
Mean of Response	-0.07166
Observations (or Sum Wgts)	674

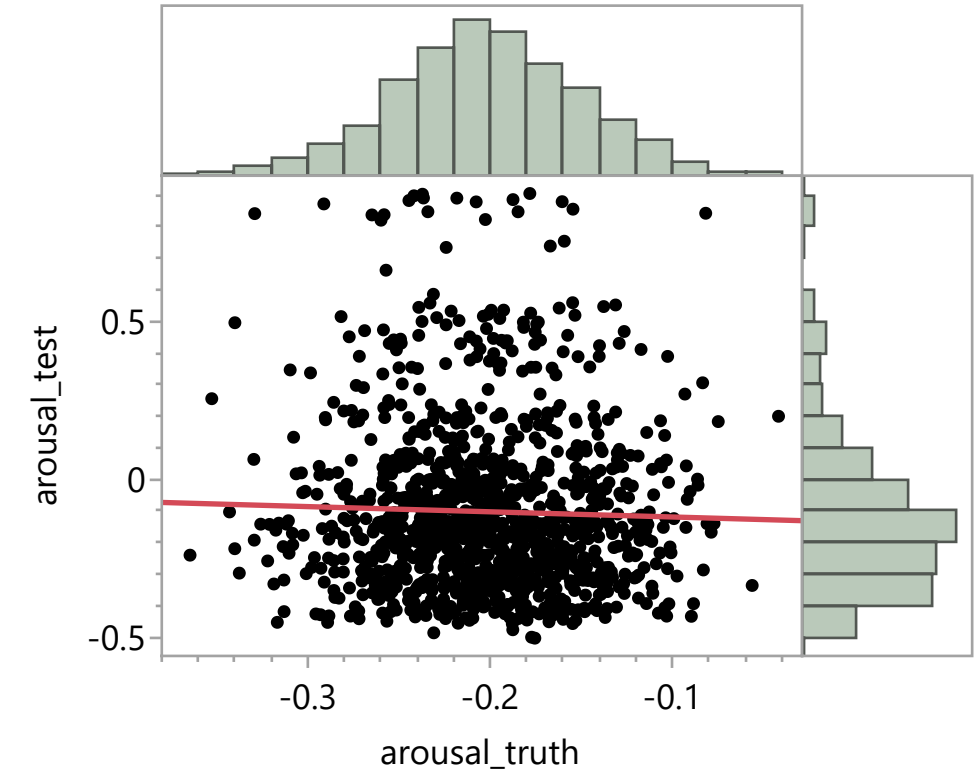
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.171523	0.171523	3.7848
Error	672	30.454420	0.045319	**Prob > F**
C. Total	673	30.625943		0.0521

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.102481	0.017839	-5.74	<.0001*
arousal_truth	0.3021813	0.155327	1.95	0.0521

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Confused



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.03104	-0.08849	0.026617	0.2913
Covariance	-0.00041			
Count	1158			

Variable	Mean	Std Dev
arousal_truth	-0.20153	0.049886
arousal_test	-0.10283	0.264034

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Confused

Linear Fit

arousal_test = -0.135933 - 0.1642739*arousal_truth

Summary of Fit

RSquare	0.000963
RSquare Adj	9.91e-5
Root Mean Square Error	0.264021
Mean of Response	-0.10283
Observations (or Sum Wgts)	1158

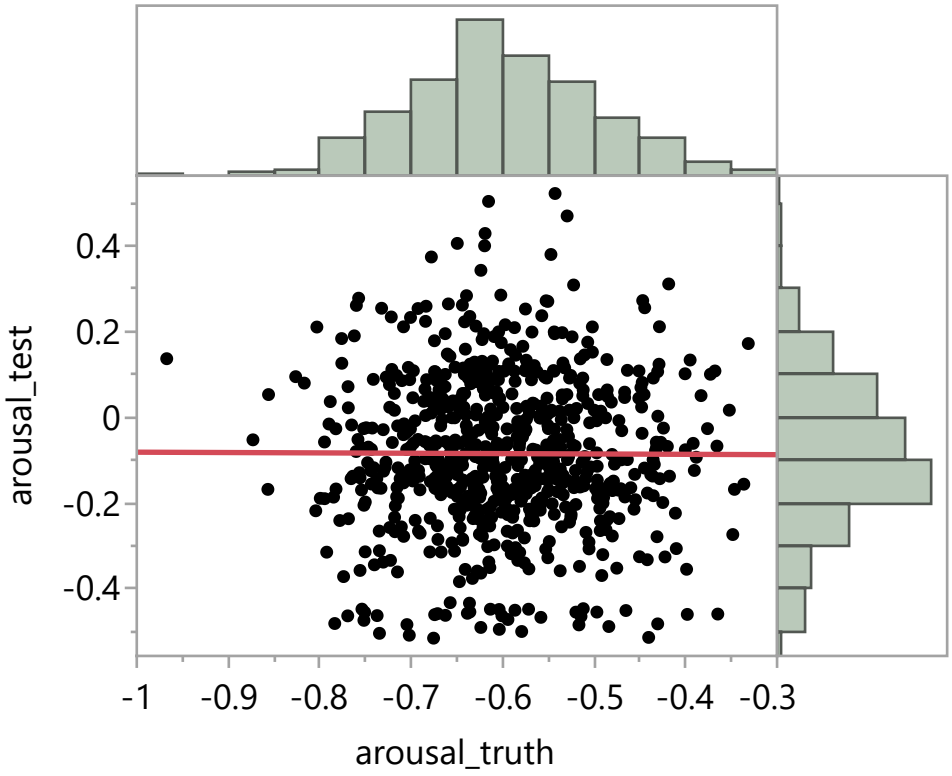
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.077701	0.077701	1.1147
Error	1156	80.581547	0.069707	**Prob > F**
C. Total	1157	80.659247		0.2913

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.135933	0.032302	-4.21	<.0001*
arousal_truth	-0.164274	0.155595	-1.06	0.2913

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Distracted



Linear Fit

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.00454	-0.07773	0.068701	0.9034
Covariance	-7.87e-5			
Count	717			

Variable	Mean	Std Dev
arousal_truth	-0.60011	0.097143
arousal_test	-0.08495	0.178351

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Distracted

Linear Fit

arousal_test = -0.089951 - 0.0083365*arousal_truth

Summary of Fit

RSquare	2.062e-5
RSquare Adj	-0.00138
Root Mean Square Error	0.178474
Mean of Response	-0.08495
Observations (or Sum Wgts)	717

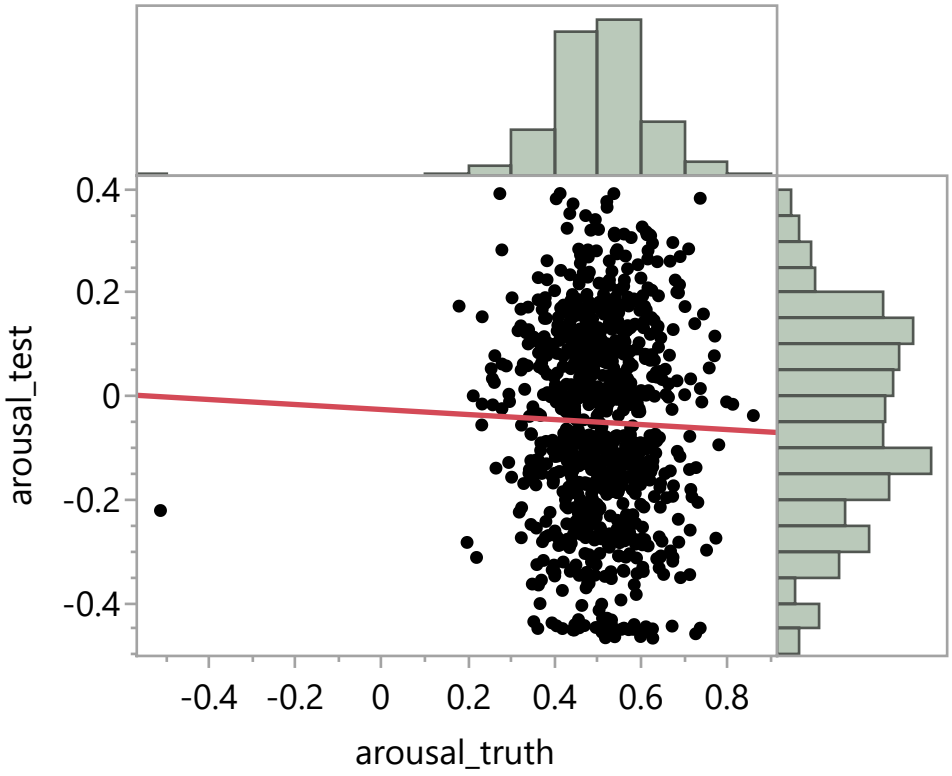
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.000470	0.000470	0.0147
Error	715	22.774778	0.031853	**Prob > F**
C. Total	716	22.775248		0.9034

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.089951	0.04174	-2.16	0.0315*
arousal_truth	-0.008336	0.06866	-0.12	0.9034

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Frustrated



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.02654	-0.09176	0.038913	0.4267
Covariance	-0.00053			
Count	899			

Variable	Mean	Std Dev
arousal_truth	0.504276	0.10467
arousal_test	-0.05033	0.191073

Bivariate Fit of arousal_test By arousal_truth class_truth_lbl=Frustrated

Linear Fit

arousal_test = -0.025896 - 0.0484467*arousal_truth

Summary of Fit

RSquare	0.000704
RSquare Adj	-0.00041
Root Mean Square Error	0.191112
Mean of Response	-0.05033
Observations (or Sum Wgts)	899

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.023091	0.023091	0.6322
Error	897	32.761734	0.036524	**Prob > F**
C. Total	898	32.784826		0.4267

Parameter Estimates

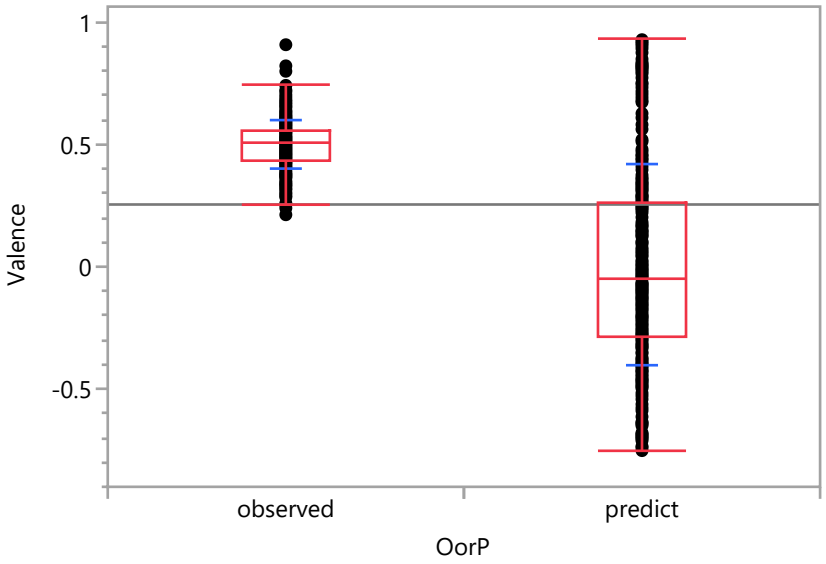
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.025896	0.031379	-0.83	0.4094
arousal_truth	-0.048447	0.060929	-0.80	0.4267

C.3 Comparison of Observed and Predicted Valence Means within Classes

The following are the results of statistical tests used to determine a difference in the means between the ground truth valence and the model generated valence scores within each category. Note that the left side of the x-axis represents the observed, or ground truth mean as identified by the tutors. The right side of the axis is the mean valence score as predicted by the model.

Correlation Check, F and T Tests by Discrete Category - Valence

Oneway Analysis of Valence By OorP class_lbl=Aha



Quantiles

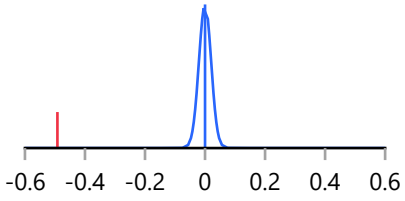
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	0.212525	0.376421	0.431891	0.504208	0.559647	0.62699	0.90759
predict	-0.75191	-0.46995	-0.28499	-0.04889	0.258028	0.70517	0.928984

t Test

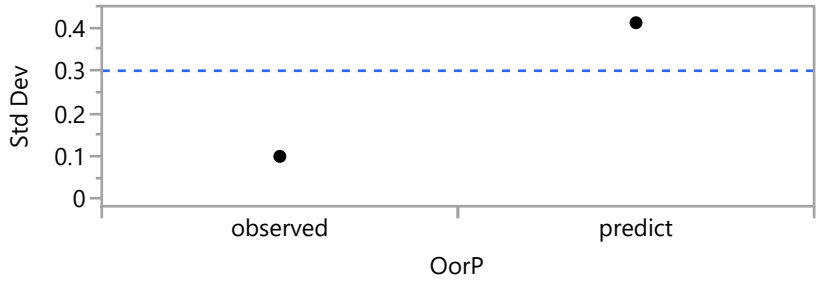
predict-observed

Assuming unequal variances

Difference	-0.49184	t Ratio	-24.2514
Std Err Dif	0.02028	DF	484.2813
Upper CL Dif	-0.45199	Prob > |t|	<.0001*
Lower CL Dif	-0.53169	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	435	0.0991371	0.0777262	0.0776678
predict	435	0.4112067	0.3316571	0.3271059

Oneway Analysis of Valence By OorP class_lbl=Aha

Tests that the Variances are Equal

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	247.3439	1	868	<.0001*
Brown-Forsythe	392.6150	1	868	<.0001*
Levene	447.9702	1	868	<.0001*
Bartlett	681.4129	1	.	<.0001*
F Test 2-sided	17.2047	434	434	<.0001*

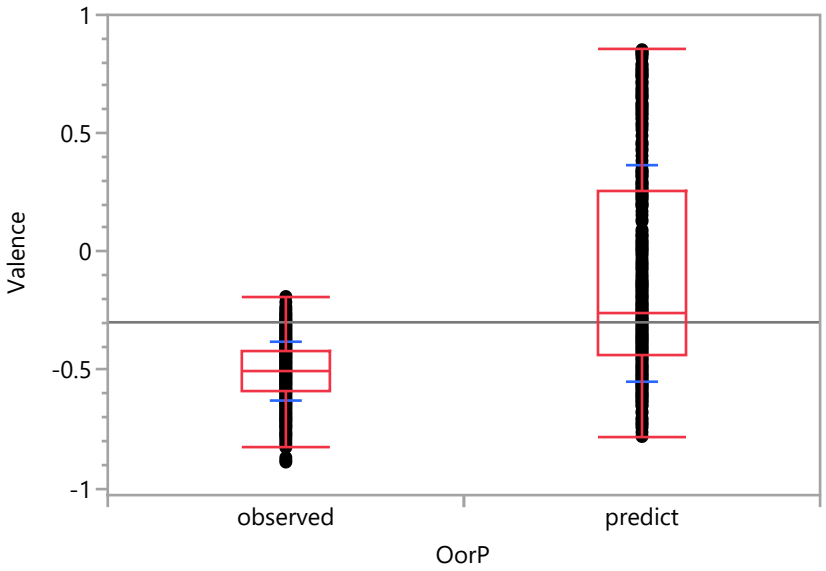
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
588.1326	1	484.28	<.0001*

t Test
24.2514

Oneway Analysis of Valence By OorP class_lbl=Anxious



Quantiles

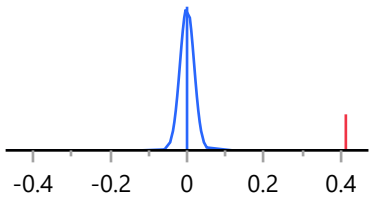
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.88977	-0.6663	-0.58849	-0.50202	-0.42083	-0.34171	-0.19004
predict	-0.78184	-0.58353	-0.43756	-0.2561	0.253247	0.678935	0.85565

t Test

predict-observed

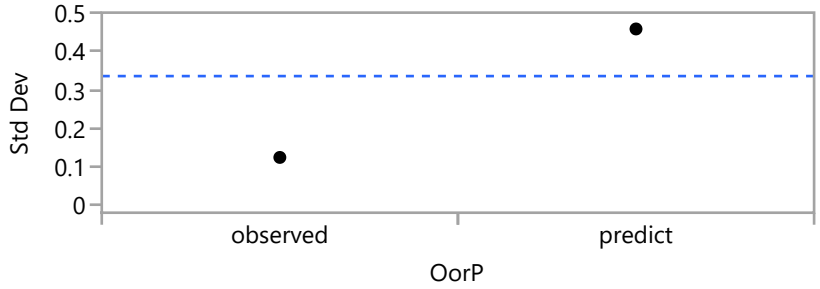
Assuming unequal variances

Difference	0.412900	t Ratio	22.60623
Std Err Dif	0.018265	DF	771.7167
Upper CL Dif	0.448755	Prob > |t|	<.0001*
Lower CL Dif	0.377046	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Oneway Analysis of Valence By OorP class_lbl=Anxious

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	674	0.1242647	0.1000477	0.1000065
predict	674	0.4576122	0.3884666	0.3682546

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	483.4661	1	1346	<.0001*
Brown-Forsythe	458.4085	1	1346	<.0001*
Levene	880.2330	1	1346	<.0001*
Bartlett	916.7635	1	.	<.0001*
F Test 2-sided	13.5612	673	673	<.0001*

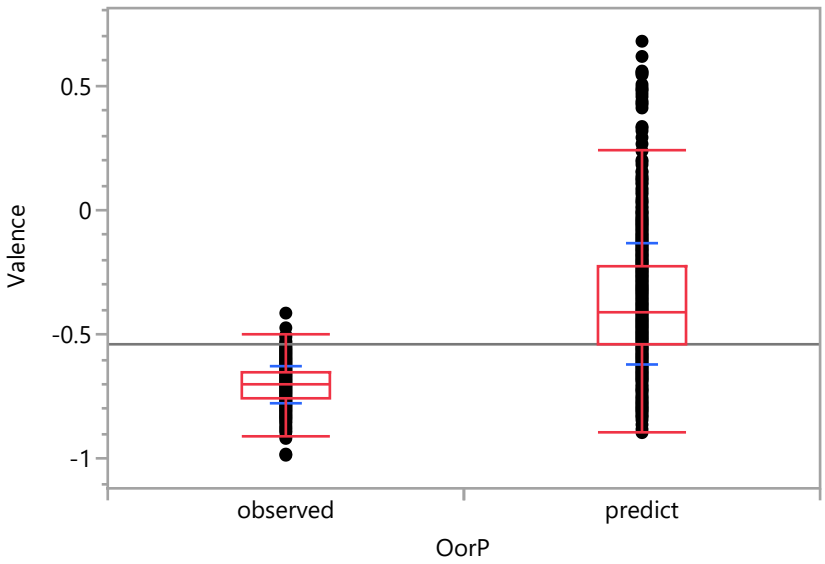
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
511.0417	1	771.72	<.0001*

t Test
22.6062

Oneway Analysis of Valence By OorP class_lbl=Confused



Oneway Analysis of Valence By OorP class_lbl=Confused

Quantiles

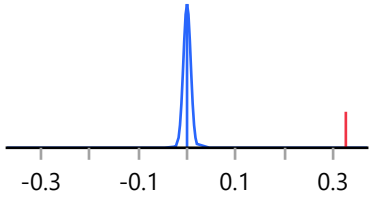
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.98535	-0.79559	-0.75462	-0.70382	-0.65008	-0.60041	-0.41413
predict	-0.89455	-0.65121	-0.54296	-0.40915	-0.22487	-0.09704	0.678673

t Test

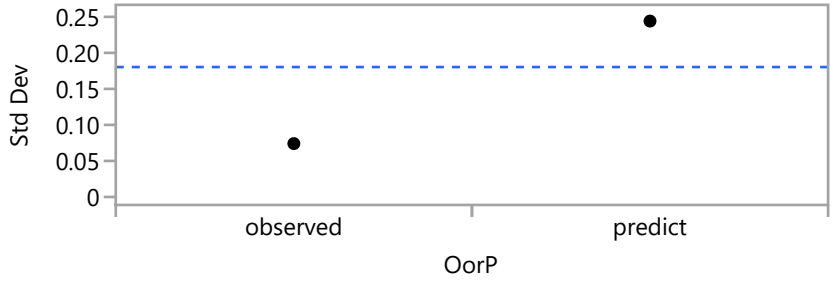
predict-observed

Assuming unequal variances

Difference	0.325138	t Ratio	43.39745
Std Err Dif	0.007492	DF	1371.416
Upper CL Dif	0.339835	Prob > |t|	<.0001*
Lower CL Dif	0.310440	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	1158	0.0745401	0.0594724	0.0594552
predict	1158	0.2438113	0.1910238	0.1891992

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	254.1897	1	2314	<.0001*
Brown-Forsythe	730.5381	1	2314	<.0001*
Levene	803.5796	1	2314	<.0001*
Bartlett	1344.4697	1	.	<.0001*
F Test 2-sided	10.6986	1157	1157	<.0001*

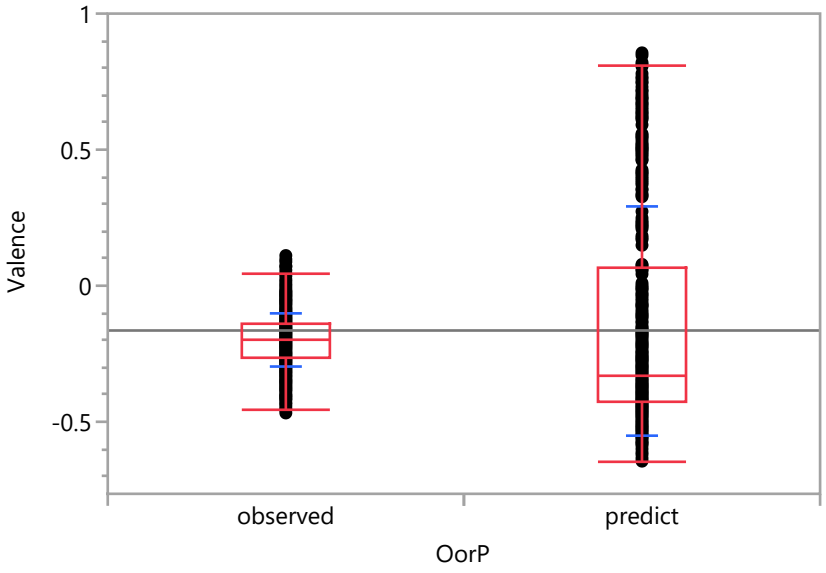
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
1883.3388	1	1371.4	<.0001*

t Test
43.3975

Oneway Analysis of Valence By OorP class_lbl=Distracted



Quantiles

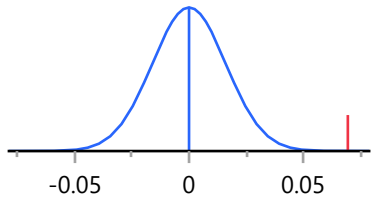
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.46745	-0.31956	-0.26438	-0.20068	-0.1369	-0.07702	0.111056
predict	-0.64448	-0.49616	-0.42738	-0.32991	0.067982	0.630275	0.855851

t Test

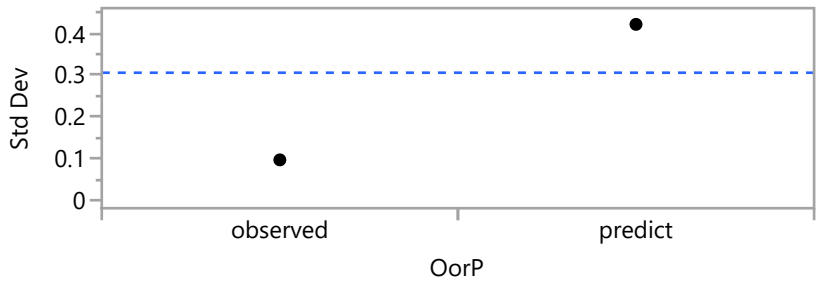
predict-observed

Assuming unequal variances

Difference	0.069450	t Ratio	4.30093
Std Err Dif	0.016148	DF	792.8401
Upper CL Dif	0.101147	Prob > |t|	<.0001*
Lower CL Dif	0.037753	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	717	0.0977092	0.0773077	0.0772966
predict	717	0.4211969	0.3508522	0.3087807

Oneway Analysis of Valence By OorP class_lbl=Distracted

Tests that the Variances are Equal

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	403.1224	1	1432	<.0001*
Brown-Forsythe	305.7726	1	1432	<.0001*
Levene	929.8233	1	1432	<.0001*
Bartlett	1173.9554	1	.	<.0001*
F Test 2-sided	18.5823	716	716	<.0001*

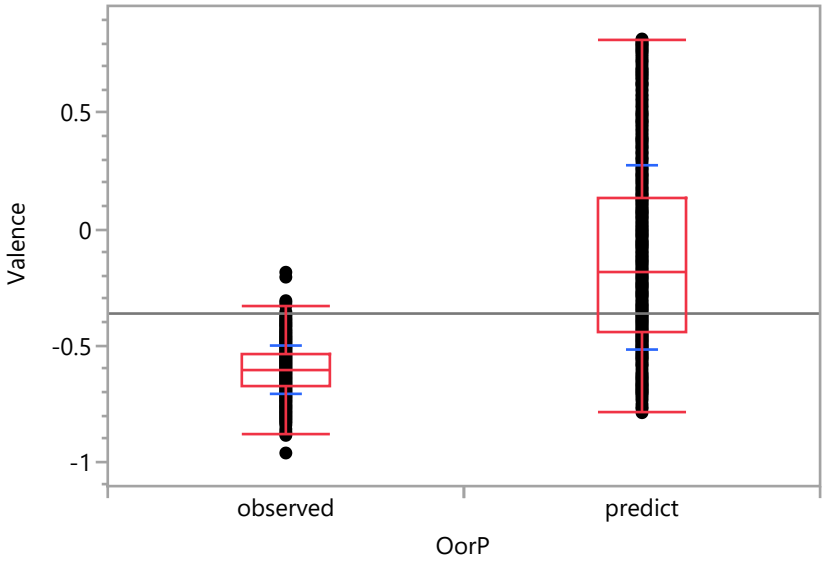
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
18.4980	1	792.84	<.0001*

t Test
4.3009

Oneway Analysis of Valence By OorP class_lbl=Frustrated



Quantiles

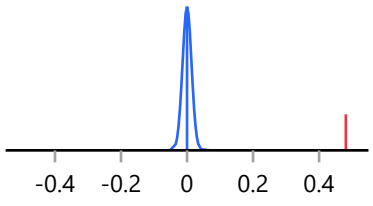
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.96206	-0.73784	-0.67291	-0.60438	-0.53473	-0.46676	-0.18512
predict	-0.78872	-0.62548	-0.43912	-0.18622	0.137862	0.464116	0.817356

t Test

predict-observed

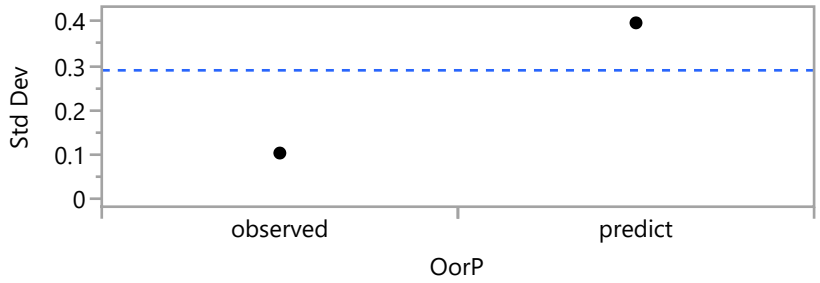
Assuming unequal variances

Difference	0.482728	t Ratio	35.34665
Std Err Dif	0.013657	DF	1020.853
Upper CL Dif	0.509527	Prob > |t|	<.0001*
Lower CL Dif	0.455929	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Oneway Analysis of Valence By OorP class_lbl=Frustrated

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	899	0.1038400	0.0827319	0.0827303
predict	899	0.3960964	0.3260598	0.3218153

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	526.3293	1	1796	<.0001*
Brown-Forsythe	837.4827	1	1796	<.0001*
Levene	978.6509	1	1796	<.0001*
Bartlett	1278.2693	1	.	<.0001*
F Test 2-sided	14.5503	898	898	<.0001*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

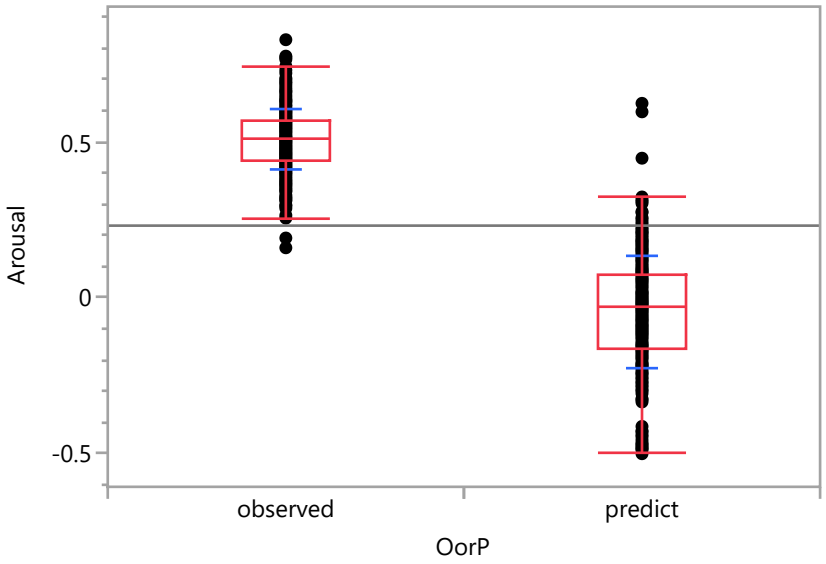
F Ratio	DFNum	DFDen	Prob > F
1249.3855	1	1020.9	<.0001*

t Test
35.3466

C.4 Comparison of Observed and Predicted Arousal Means within Classes

The following are the results of statistical tests used to determine a difference in the means between the ground truth arousal and the model generated arousal scores within each category. Note that the left side of the x-axis represents the mean observed, or ground truth arousal as identified by the tutors. The right side of the axis is the mean arousal score as predicted by the model.

Oneway Analysis of Arousal By OorP class_lbl=Aha



Quantiles

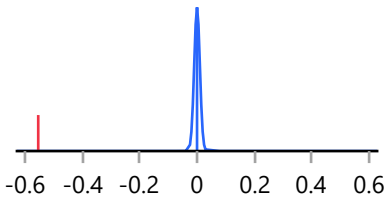
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	0.160127	0.390312	0.442585	0.510274	0.570753	0.63175	0.828057
predict	-0.50098	-0.25935	-0.1658	-0.0297	0.073132	0.177715	0.623798

t Test

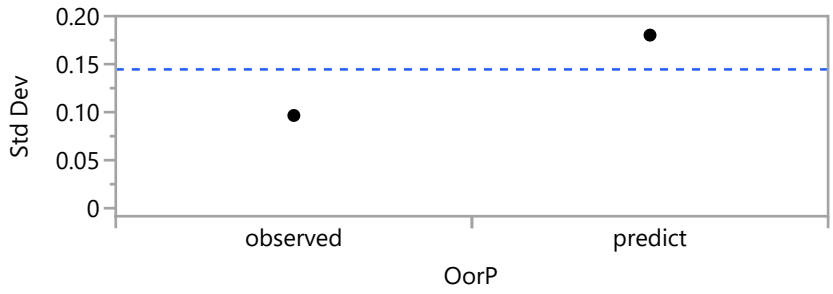
predict-observed

Assuming unequal variances

Difference	-0.55465	t Ratio	-56.5088
Std Err Dif	0.00982	DF	665.5565
Upper CL Dif	-0.53537	Prob > |t|	<.0001*
Lower CL Dif	-0.57392	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	435	0.0969397	0.0757698	0.0757551
predict	435	0.1803049	0.1376148	0.1367793

Oneway Analysis of Arousal By OorP class_lbl=Aha

Tests that the Variances are Equal

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	76.3567	1	868	<.0001*
Brown-Forsythe	91.6503	1	868	<.0001*
Levene	96.8931	1	868	<.0001*
Bartlett	157.2101	1	.	<.0001*
F Test 2-sided	3.4595	434	434	<.0001*

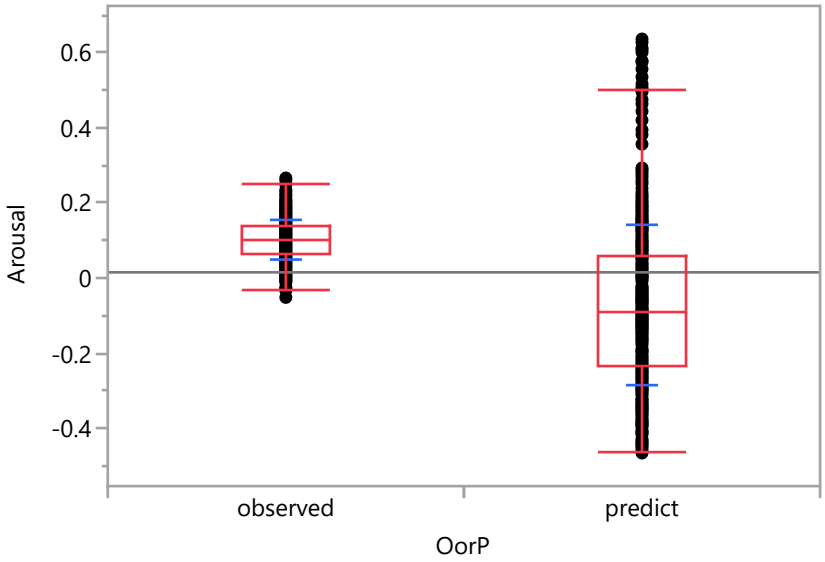
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
3193.2456	1	665.56	<.0001*

t Test
56.5088

Oneway Analysis of Arousal By OorP class_lbl=Anxious



Quantiles

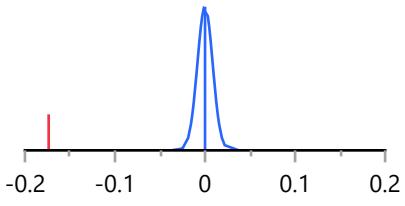
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.05154	0.034324	0.064565	0.100992	0.138517	0.172934	0.266348
predict	-0.46555	-0.3492	-0.23679	-0.0886	0.060012	0.168572	0.636067

t Test

predict-observed

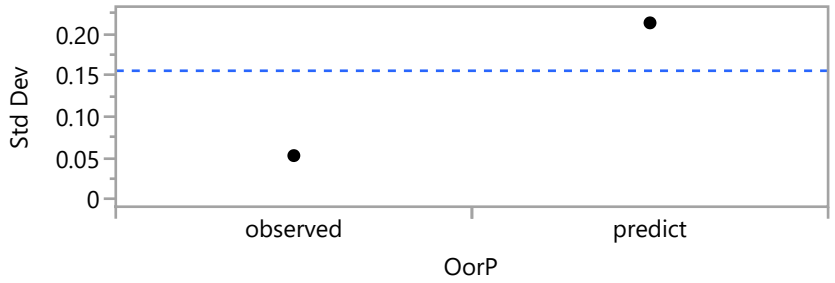
Assuming unequal variances

Difference	-0.17366	t Ratio	-20.5144
Std Err Dif	0.00847	DF	755.2456
Upper CL Dif	-0.15704	Prob > |t|	<.0001*
Lower CL Dif	-0.19027	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*



Oneway Analysis of Arousal By OorP class_lbl=Anxious

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	674	0.0528307	0.0422446	0.0422329
predict	674	0.2133228	0.1626724	0.1619536

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	213.7321	1	1346	<.0001*
Brown-Forsythe	470.5095	1	1346	<.0001*
Levene	488.5322	1	1346	<.0001*
Bartlett	1025.0158	1	.	<.0001*
F Test 2-sided	16.3043	673	673	<.0001*

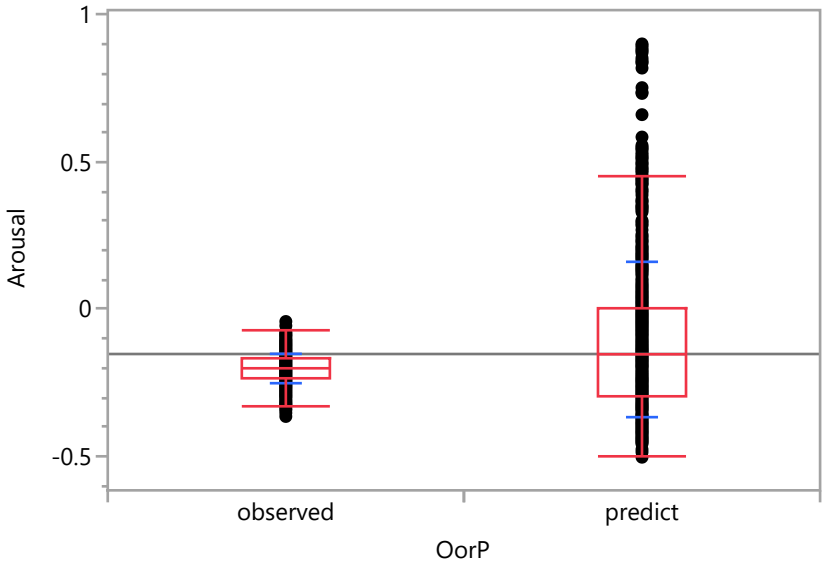
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
420.8408	1	755.25	<.0001*

t Test
20.5144

Oneway Analysis of Arousal By OorP class_lbl=Confused



Oneway Analysis of Arousal By OorP class_lbl=Confused

Quantiles

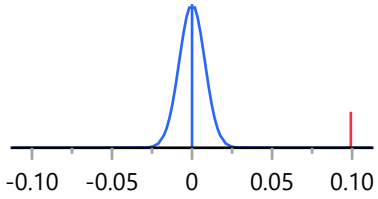
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.36449	-0.26327	-0.23475	-0.20265	-0.16712	-0.13672	-0.04165
predict	-0.503	-0.37658	-0.29454	-0.15485	0.005184	0.235207	0.90426

t Test

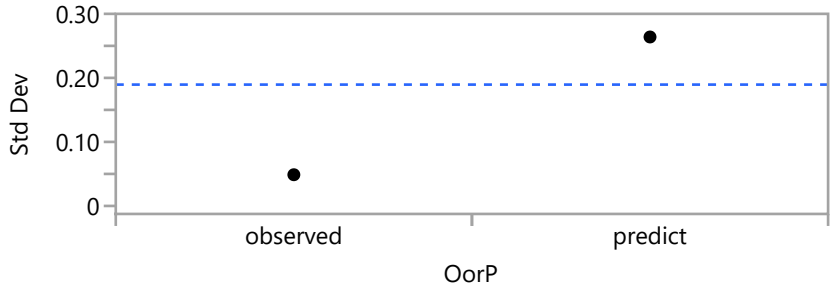
predict-observed

Assuming unequal variances

Difference	0.098697	t Ratio	12.4992
Std Err Dif	0.007896	DF	1239.498
Upper CL Dif	0.114189	Prob > |t|	<.0001*
Lower CL Dif	0.083206	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	1158	0.0498858	0.0397026	0.0396897
predict	1158	0.2640343	0.1990972	0.1924100

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	261.5108	1	2314	<.0001*
Brown-Forsythe	744.3919	1	2314	<.0001*
Levene	950.5597	1	2314	<.0001*
Bartlett	2332.1280	1	.	<.0001*
F Test 2-sided	28.0134	1157	1157	<.0001*

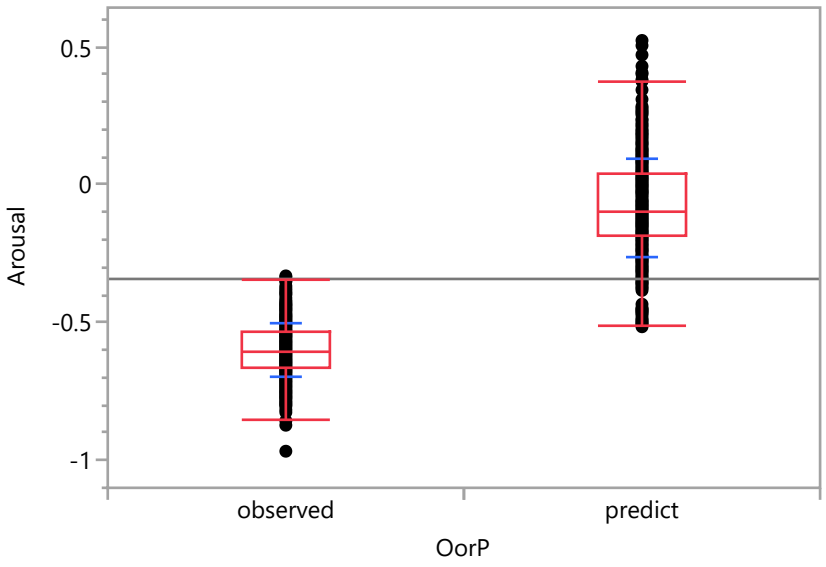
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
156.2300	1	1239.5	<.0001*

t Test
12.4992

Oneway Analysis of Arousal By OorP class_lbl=Distracted



Quantiles

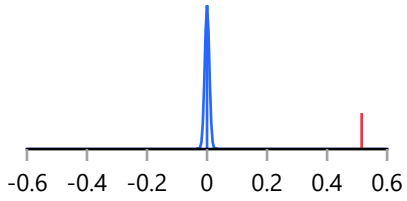
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.96725	-0.72446	-0.66618	-0.60583	-0.53662	-0.47248	-0.33147
predict	-0.51585	-0.31185	-0.18728	-0.09634	0.038331	0.126083	0.522372

t Test

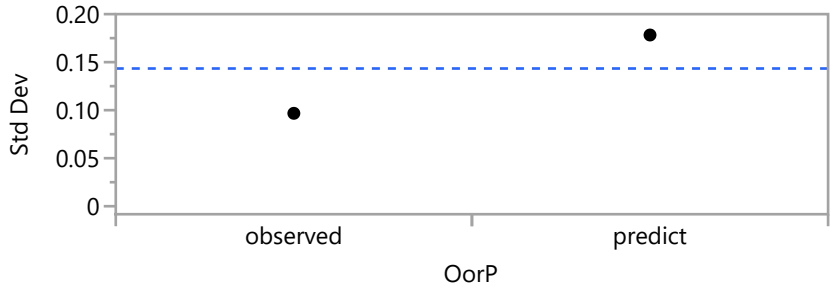
predict-observed

Assuming unequal variances

Difference	0.515167	t Ratio	67.92318
Std Err Dif	0.007585	DF	1106.464
Upper CL Dif	0.530049	Prob > |t|	<.0001*
Lower CL Dif	0.500285	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	717	0.0971429	0.0774314	0.0772820
predict	717	0.1783508	0.1399508	0.1397165

Oneway Analysis of Arousal By OorP class_lbl=Distracted

Tests that the Variances are Equal

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	146.2244	1	1432	<.0001*
Brown-Forsythe	176.0077	1	1432	<.0001*
Levene	179.3251	1	1432	<.0001*
Bartlett	249.3100	1	.	<.0001*
F Test 2-sided	3.3708	716	716	<.0001*

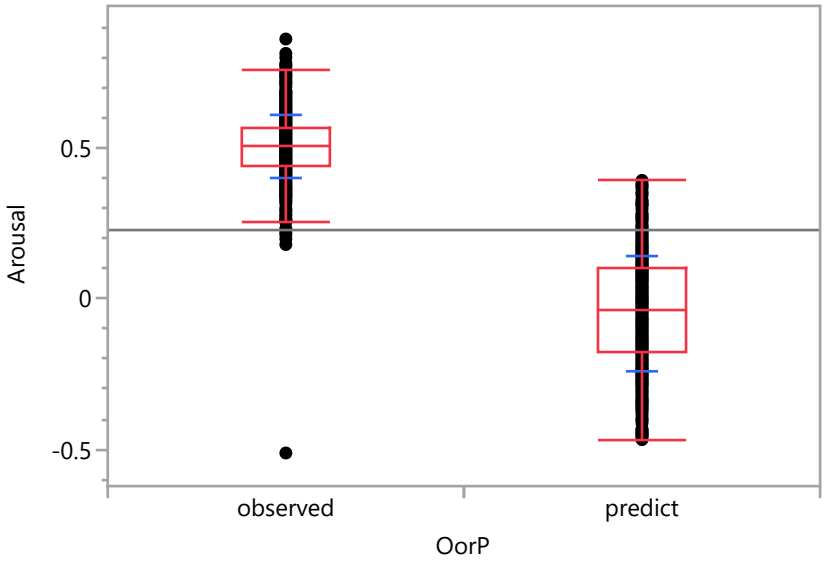
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
4613.5582	1	1106.5	<.0001*

t Test
67.9232

Oneway Analysis of Arousal By OorP class_lbl=Frustrated



Quantiles

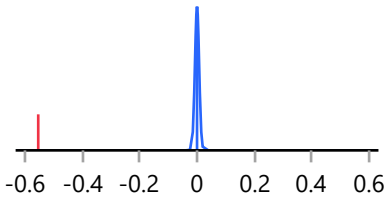
Level	Minimum	10%	25%	Median	75%	90%	Maximum
observed	-0.5124	0.376896	0.441563	0.505047	0.568886	0.626961	0.860536
predict	-0.46801	-0.30885	-0.17637	-0.03757	0.103067	0.179739	0.391853

t Test

predict-observed

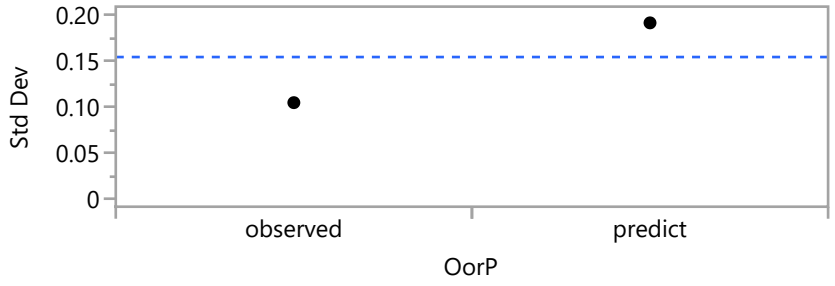
Assuming unequal variances

Difference	-0.55460	t Ratio	-76.3268
Std Err Dif	0.00727	DF	1392.434
Upper CL Dif	-0.54035	Prob > |t|	<.0001*
Lower CL Dif	-0.56886	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*



Oneway Analysis of Arousal By OorP class_lbl=Frustrated

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
observed	899	0.1046702	0.0781716	0.0781662
predict	899	0.1910725	0.1595653	0.1594446

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	180.2582	1	1796	<.0001*
Brown-Forsythe	369.7904	1	1796	<.0001*
Levene	375.5823	1	1796	<.0001*
Bartlett	307.1672	1	.	<.0001*
F Test 2-sided	3.3323	898	898	<.0001*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
5825.7781	1	1392.4	<.0001*

t Test
76.3268

Vita

Kriss Gabourel was born and raised in Southern California. Upon graduating from Lake Elsinore High School, she moved to Tallahassee, Florida to attend Florida Agricultural and Mechanical University. She graduated in 1991 with a Bachelor of Science degree in Computer Information Systems. After graduation, she remained in Tallahassee and worked as a project analyst for the State of Florida. Soon after getting married, she moved to Knoxville, Tennessee. After a brief stint in IT, she went a different route and began working at Roane State Community College. It was her first job at a postsecondary institution and it ignited a passion in her for higher education. She was eager to combine her IT skills and love of higher education to truly make a difference in the lives of students. In 2013 she began working at University of Tennessee and began pursuit of her Master's degree. She graduated with a Masters of Science in Business Analytics in 2017. The following year, she received a fellowship from the University of Tennessee Bredesen Center and returned to school full time to work on her PhD in Data Science Engineering. While pursuing her PhD, she was a research assistant at UT's Postsecondary Education Research Center, where she was pleased to hone her researching skills while tackling fascinating topics surrounding higher education policy and practice.