The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto

ECE496Y Design Project Course
Group Final Report

F.A.C.E. - Facial and Audio-based Classification of Emotion

Team 201913, Section 4

Under the supervision of Tarek Abdelrahman

Under the administration of Phil Anderson

March 31, 2020

Meghan Muldoon	meghan.muldoon@mail.utoronto.ca
Olivia Roscoe	olivia.roscoe@mail.utoronto.ca
Sofia Tijanic	sofia.tijanic@mail.utoronto.ca
Aleksei Wan	aleksei.wan@mail.utoronto.ca

Attribution Table

Section	Student Names			
	Meghan Muldoon	Olivia Roscoe	Sofia Tijanic	Aleksei Wan
Executive Summary			ET	RD
Group Highlights & Individual Contributions	RD	RD	MR	MR
Acknowledgements	RD			
Background & Motivation	RS, RD	RS, ET	RS, ET	RS, ET
Project Goal		RD	ET	
Project Requirements	ET		RD	
System-level Overview	ET		RD	
Module-Level Design		ET	MR	RS, RD
Module-Level Description	MR	MR	ET	RD
Final Design Decisions	ET	RD	ET	RS, ET
Verification Matrix	RD		MR	
Final Test Results		ET	RD	ET
Conclusion		ET	RD	
All	FP	FP	FP, CM	FP

Signatures

By signing below, you verify that you have read the attribution table and agree that it accurately reflects your contribution to this document.

Name: Meghan Muldoon Signature: M. Muldoon Date: 03/31/2020

Name: Olivia Roscoe Signature: ORoscoe Date: 03/31/2020

Name: Sofia Tijanic Signature: Sofia Tijanic Date: 03/31/2020

Name: Aleksei Wan Signature: Date: 03/31/2020

Voluntary Document Release Consent Form

To all ECE496 students:

To better help future students, we would like to provide examples that are drawn from excerpts of past student reports. The examples will be used to illustrate general communication principles as well as how the document guidelines can be applied to a variety of categories of design projects (e.g. electronics, computer, software, networking, research).

Any material chosen for the examples will be altered so that all names are removed. In addition, where possible, much of the technical details will also be removed so that the structure or presentation style are highlighted rather than the original technical content. These examples will be made available to students on the course website, and in general may be accessible by the public. The original reports will not be released but will be accessible only to the course instructors and administrative staff.

Participation is completely voluntary and students may refuse to participate or may withdraw their permission at any time. Reports will only be used with the signed consent of all team members. Participating will have no influence on the grading of your work and there is no penalty for not taking part.

If your group agrees to take part, please have all members sign the bottom of this form. The original completed and signed form should be included in the hardcopies of the final report.

Sincerely,
Khoman Phang
Phil Anderson
ECE496Y Course Coordinators

Consent Statement

We verify that we have read the above letter and are giving permission for the ECE496 course coordinator to use our reports as outlined above.

Team #: 13 Project Title: F.A.C.E - Facial and Audio-based Classification of Emotion

Supervisor: Tarek Abdelrahman Administrator: Phil Anderson

Name: Meghan Muldoon Signature: M. Muldoon Date: 03/31/2020

Name: Olivia Roscoe Signature: ORoscoe Date: 03/31/2020

Name: Sofia Tijanic Signature: Sofia Tijanic Date: 03/31/2020

Name: Aleksei Wan Signature: Date: 03/31/2020

Executive Summary (Author: A. Wan)

Autism Spectrum Disorder (ASD) is a severe developmental disability, which involves "impairments in social interaction… and verbal and nonverbal communication" [1]. Our team has seen first hand the challenges that these children experience in communicating. With this understanding, the team decided on the goal of building an application that will act as a learning tool for children with autism who have difficulty interpreting verbal and nonverbal social cues. This application is a web application designed to analyze video and audio data and produce results in an easy-to-understand interface, allowing children to practice and better comprehend the world around them.

After outlining the project goal, key functional requirements, objectives, and constraints (FOCs), we established an appropriate scope for the project and refined what would be expected from the final solution. This includes expectations on what will be classified and requirements for compliance with applicable laws and regulations. These FOCs also include validation/acceptance tests which define precisely how they would be verified.

Over the course of several months, the team designed and executed the components of the system, and created an application that is able to execute its intended function. The final design consists of five primary components, which were built modularly and integrated into a single pipeline. The components consist of a user interface, data pre-processing, audio neural network, image-based classification, and a gradient boosting decision tree. The technical challenges of this project led to important design decisions, including type of data input and machine learning methods used. The final product is a functional application that is targeted towards children with autism as a learning tool. The application is able to accept video files as input and provide an emotion prediction with an accuracy of approximately 71%.

The validation tests that the team set out to achieve show the success of the final application. These tests range from quantitative numerical results to interactive, useability features, and test all components of our system. Overall, the team has observed that the functionality and purpose of the project was achieved. We recognize that some of the technical challenges in the nature of this project and specifically with audio data analysis and classification led to final accuracies slightly below our intended numbers. However, we see this as a learning process, and opportunity for future steps to be taken. The team sees an exciting potential future for this application, with essential next steps to improve technical and useability components of the application. We believe that there is a need for these kinds of educational tools in society, and that our application could fill this gap.

Group Highlights and Individual Contributions

Group Highlights (Author: M. Muldoon)

The team has achieved several notable accomplishments throughout the completion of the project; some highlights include major decisions that were made, the implementation of key technical components, and many new findings that the group discovered along the way.

The audio data analysis and classification done in this project proved to be one of the largest and most challenging components. As a result, many of the design decisions made for this component came from trial and error, as well as the team's best educated estimate. Through research, the team made the decision to use Mel-Frequency Cepstral Coefficient (MFCC) analysis to perform audio-feature extraction. This step was done in the preprocessing stage in order to prepare video data to be analyzed and classified. The team completed a data preprocessing script, which extracts desired features from video clips, creating spectrogram images for audio classification and frame images for video classification.

In addition to this, the team created an audio neural network from scratch which was trained to generate an emotion classification based on a spectrogram image generated from speech. This component involved iterative design and trial of various neural network architectures, as well as parameters. The team had to think beyond traditional machine learning approaches, and use visual classification networks on audio data. The achievement of this component stemmed from a cross between creativity and technical knowledge, which the team sees as a highlight and important design choice throughout the project.

In combining the AWS classification with our custom neural network results, the team took another notable, creative approach: to combine the results using a gradient boosting decision tree. This component is trained to look for patterns in the results, and come to the best final classification decision. This combination technique was chosen over traditional weighted analysis as a more technically challenging, but accuracy boosting result.

Finally, all of these components were successfully combined into one classification pipeline that takes an input video, pre-processes it, classifies the components and combines the results into one final emotion prediction. The team opted to design two separate user interfaces to display this project; one as an educational tool for children with autism, and the other as a technical demonstration of the classification pipeline. In combining the user interfaces with the back-end pipeline, the team was able to complete the final application design.

Individual Contributions (Author: O.Roscoe)

Meghan contributed to the project in areas related mostly to audio feature extraction for emotion recognition. Meghan took on the challenge of creating the pre-processing script for input video clips. With help from Aleksei this script was able to take input video clips and create facial image and spectrogram frames which could be fed directly to the separate neutral networks. This required performing research on methods of processing audio data as well as research on technical implementation of these methods using available Python libraries.

Olivia's main contribution was building and training the neural network to classify the audio data. Several different architectures were tested to try and achieve the best possible validation accuracy for the spectral image data used as input. This involved building multiple networks, testing different methods of separating the data into training, validation and testing, tuning the hyperparameters of the network throughout training sessions and merging multiple datasets. She analyzed the results of each training session in an attempt to improve the validation accuracy, and though they did not achieve the final accuracy they wanted, significant improvement was made from the initial model. Olivia also developed baseline models for the image classification to compare to other solutions.

Sofia worked on the two user interface components of the project, as well as the classification pipeline that integrated all components together and tied them into the front end. The user interface designed for children with autism was heavily influenced by research [2] on the design principles for this type of user. The secondary user interface was developed as a way of demonstrating the classification pipeline & showing the various results achieved throughout, additional data was extracted from various modules, sorted, and then displayed in a real-time. In addition to the front-end, Sofia created the pipeline using all of the modular components built by Meghan, Olivia and Aleksei. The pipeline connects all of the required outputs and inputs. This pipeline is integrated with both user interfaces, so as to be triggered by a user.

Aleksei's primary contributions were focussed on the image classification and the final prediction optimization. For image classification, Aleksei identified AWS Rekognition, and wrote the scripts necessary to test it, compare it to a baseline prepared by Olivia, and implement it as our final image-based solution. Additionally, Aleksei helped across other dimensions of the project (e.g. assisting Olivia with tuning the audio network). Finally, Aleksei created the solution for merging the audio-based and image-based prediction components. This included testing a baseline and advanced solutions. This resulted in Aleksei implementing LightGBM, a gradient-boosted decision tree, along with integration with the upstream predictions from the image-based and audio-based networks, and probabilistically sweeping the parameter space during training. This implementation became the team's final classification component.

Acknowledgements (Author: M. Muldoon)

The group would like to acknowledge people and organizations without which this project would not have been possible. Firstly we would like to thank Professor Tarek Abdelrahman for supervising our team throughout the process, aiding us in the development of the project scope and for providing us with access to computational resources necessary to the completion of the project. In addition we would also like to thank our administrator Professor Anderson for encouraging our team and providing helpful feedback throughout the report and presentation forming process. Finally we would like to acknowledge the SMART Lab in the Ryerson University Department of Psychology and the Department of Computer Science at Cheyney University of Pennsylvania for providing public domain audio-video emotional databases that were integral to the completion of the project.

We would also like to thank the University of Toronto, Faculty of Engineering and Electrical and Computer Engineering program for enabling us to gain the relevant tools, resources and knowledge to embark on this project. Likewise, we would like to thank our peers for supporting and inspiring us throughout the course of this year.

Table of Contents

Introduction

Background and Motivation (Author: M. Muldoon)

Autism Spectrum Disorder (ASD) is a developmental disability characterised by impairments in social interaction [1]. As a team we have had experience working with children diagnosed with ASD and have seen first hand the difficulties that these children encounter while communicating with teachers and peers. In response to this, the goal of our project is to build an application that will act as a learning tool for children on the autism spectrum who struggle with verbal and nonverbal social cues.

Children with ASD often have an inherent difficulty recognizing facial emotions, as well as the vocal social cues used in non-literal speech such as sarcasm, humour, and irony [3]. Though these impairments are characteristic to the Autism Spectrum Disorder, evidence suggests that children with ASD have the ability to improve upon their understanding of verbal and nonverbal social cues through practice and repeated exposure [4]. Our project will aid in understanding and teaching essential verbal and visual cues to children, and will be designed to fulfill this purpose.

Our project seeks to investigate the challenging area of interpreting social cues and emotion from speech and facial expressions. Specifically, analyzing human speech is a process that remains highly unsolved, while facial recognition and emotion identification from still images is an area with far more research. The project will investigate both identifying emotion from speech, and combining this with facial emotion recognition to build a unique and accurate tool. Using insights from this investigation we aim to bridge the gap between members of society who can effortlessly understand their peers, and those who must constantly strive to identify social cues. The project focus will be in the specialized cases of children with autism.

Project Goal (Author: O.Roscoe)

The goal of this project is to develop a web application that can identify emotions and social cues, from audio and visual data. The user will provide a video stream, either pre-recorded or live, of social cues exhibited by an individual that they would like characterized. The program will then process the collected audio and video data, and the analysis results will then be shown to the user in an intuitive interface, so the user can practice associating the social cue with what they saw. If the user would like more information, a report of the analysis will also be available.

Project Requirements (Author: S. Tijanic)

Table 1: Project Functional Requirements

ID	Functional Requirements	Description
1.1	Output correctly classified emotions with minimum 60% test accuracy, and an objective 80% accuracy.	Solution is able to provide user with a correct classification of emotion
1.2	Accept video and audio data as input.	Solution is based on input in the form of video and audio data.
1.3	Achieve minimum 70% accuracy in emotion classification from image classification.	Facial emotion component of solution is able to provide stand-alone correct classification of emotion with 70% accuracy.
1.4	Achieve minimum 60% accuracy in emotion classification from audio classification.	Audio emotion component of solution is able to provide stand-alone correct classification of emotion with 60% accuracy.

Table 2: Project Objectives

ID	Objectives	Description
2.1	Correctly classify between 3-5 human emotions.	Use video and audio input to classify emotions based on the input data provided for emotion classification.
2.2	Minimize time required to find and classify a solution to be within 10 seconds.	Lower time required to classify input will result in a better and more easily interactive user experience [5].

Table 3: Project Constraints

ID	Constraints	Description
3.1	Data used in any phase of the project shall comply with germane Canadian data privacy legislation.	The Personal Information Protection and Electronic Documents Act (PIPEDA) dictates how data can be obtained and used by individuals and businesses. All data obtained shall comply with the principles outlined in this Act.

Final Design

System-Level Overview (Author: S. Tijanic)

The final design for this project consists of a modular system that can take an mp4 file as input, run it through a machine learning classification pipeline, and return a final emotion prediction. This system is accessible as an application with a user-interface where the classification pipeline is triggered from one click. This final design closely follows the proposed solution, with modifications and improvements made where needed. The following sections will serve as an in-depth look into each of the components of our system, how they are designed and why certain design decisions were made.

The core functionality of this application can be divided into four main components (see Figure 1 below). The first component is user interface and data capture, which involves an mp4 video file being uploaded to the system via a user interface, and the file then being run through the classification pipeline. The second module is the pre-processing component of the system. In pre-processing, the input file is separated into audio and image components, and prepared for further analysis. After this, component 3 analyzes the audio and visual pre-processed data using machine learning networks, providing classification predictions. These predictions are finally fed into the fourth component, classification refinement, consisting of a gradient boosting decision tree (LightGBM). Once the decision tree outputs a final prediction, the pipeline returns the result back to the first component where they are displayed in the user interface.

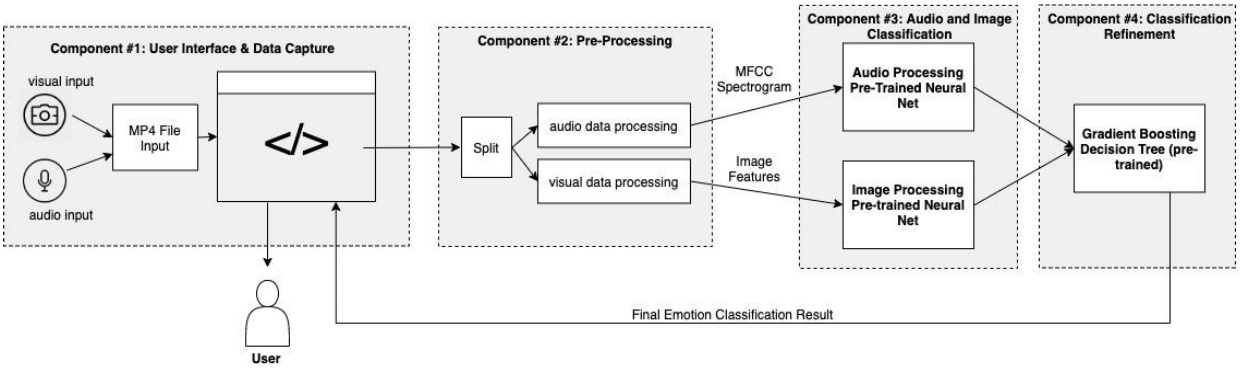

Component #1: User Interface & Data Capture
visual input
audio input
MP4 File Input
</>
User
Component #2: Pre-Processing
Split
audio data processing
visual data processing
MFCC Spectrogram
Image Features
Component #3: Audio and Image Classification
Audio Processing Pre-Trained Neural Net
Image Processing Pre-trained Neural Net
Component #4: Classification Refinement
Gradient Boosting Decision Tree (pre-trained)
Final Emotion Classification Result

Figure 1 - System Overview Diagram

Module-Level Design (Author: A. Wan)

The following tables describe the modules from Figure 1 in further detail.

Table 5: Component #1 - User Interface & Data Capture

User Interface	Input(s)	- *.mp4 video from a user, containing one human speaking in one continuous emotion - Final emotional classification (from Component # 4: Classification Refinement)
	Output	- *.mp4 video (to Preprocessing via an API call) - Emotion Classification (to User)
	Function	This module is a convenient interface for the user, sitting between them and the processing/classification modules.

Table 6: Component #2 - Preprocessing

Split	Input(s)	- *.mp4 video from UI
	Output	- *.wav file of extracted audio - *.jpg files of individual frames
	Function	This module separates the audio in the file, and samples frames as still images for processing
Audio Data Processing	Input(s)	- *.wav file
	Output	- *.jpg MFC Coefficient (MFCC) spectral image
	Function	This module performs Mel-Frequency Cepstral (MFC) analysis on the input *.wav file to produce a MFCC spectral image that can be run through a CNN
Visual Data Processing	Input(s)	- *.jpg files of individual frames
	Output	- Cropped *.jpg files of individual frames
	Function	This module removes noise by cropping the background area of the image and focusing the analysis on the subject.

Table 7: Component #3 - Audio & Image Classification

Audio	Input(s)	- MFCC spectrogram (*.jpg)

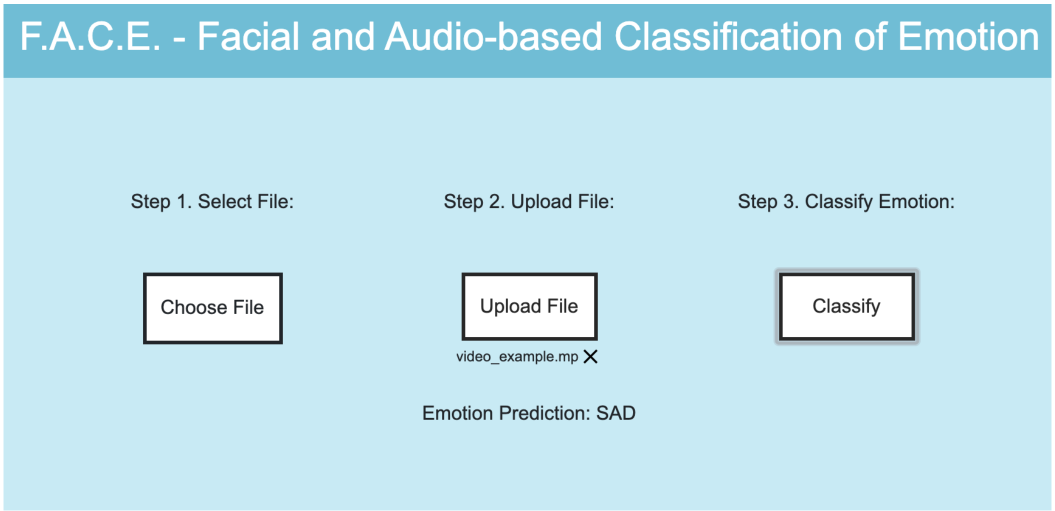
Neural Network	Output	- Emotional classification based on audio data
	Function	This network classifies the emotion seen in the spectrogram to one of the predefined classes.
Image Neural Network	Input(s)	- Cropped image frames (*.jpg)
	Output	- Emotional classification based on image data
	Function	This network classifies the emotion seen in the image to one of the predefined classes.

Table 8: Component #4 - Classification Refinement

Gradient Boosting Decision Tree (GBDT)	Input(s)	- Predicted emotion from audio network - Predicted emotion from image network
	Output	- Final, unified emotion prediction
	Function	This module uses GBDTs, a widely-used algorithm to quickly and efficiently train decision tree-like structures that produce accurate outcomes [6]. It will be trained to efficiently fuse the audio and image data to produce a final classification.

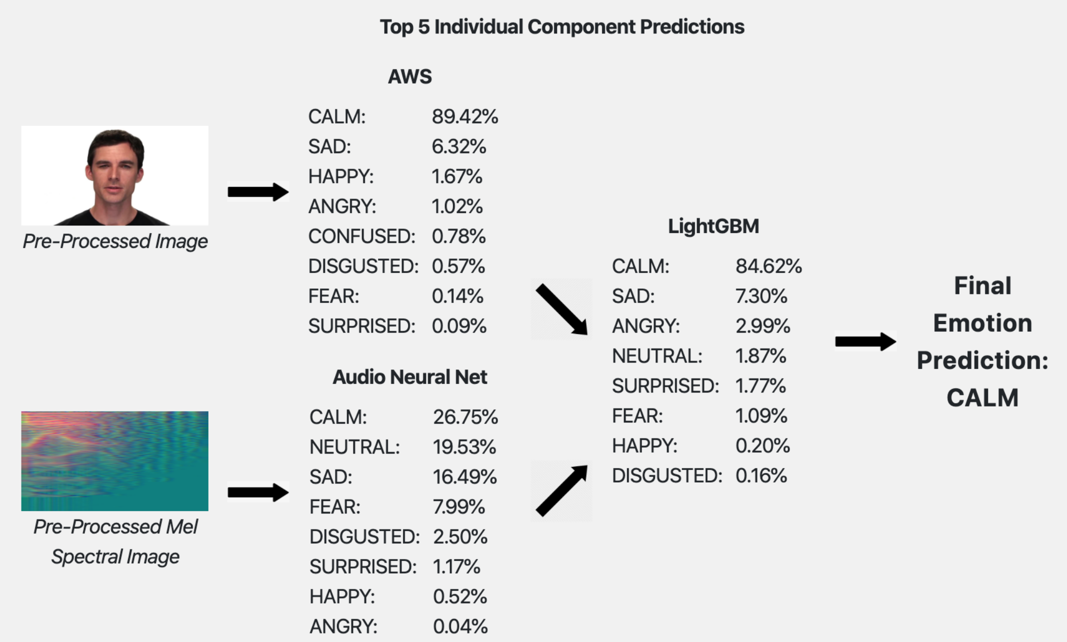
Module-Level Description (Author: A. Wan)

The first component of our system - the user interface and data capture - is the way that our application interacts with the world. As described in our project motivation, this application was inspired by a gap in educational tools for children with autism. Likewise, this project poses many technical challenges, and investigates unexplored problems. As a result, we have developed two separate user interface designs for two different use cases: (1) an educational tool for children with autism, and (2) a demonstration of our classification system and what results we were able to create. The first interface was designed using research on users with autism [2]. The goal was to follow basic design principles outlined in the research, such as simplicity, and fluidity of information presentation. The second interface was designed with the goal being to show intermediary data from the various components of our system. This includes the pre-processed data, and predictions made by our audio and visual components. Both user interfaces were developed using a Javascript React framework, with a Javascript Node server, and can be seen in the Figures 2 and 3 below, as well as in Appendix C Figures C10 and C11.



Emotion Prediction: SAD

Figure 2: User Interface #1 - Educational Tool for Children with Autism


Top 5 Individual Component Predictions
Pre-Processed Image
Pre-Processed Mel Spectral Image
Final Emotion Prediction: CALM

AWS

CALM:	89.42%
SAD:	6.32%
HAPPY:	1.67%
ANGRY:	1.02%
CONFUSED:	0.78%
DISGUSTED:	0.57%
FEAR:	0.14%
SURPRISED:	0.09%

Audio Neural Net

CALM:	26.75%
NEUTRAL:	19.53%
SAD:	16.49%
FEAR:	7.99%
DISGUSTED:	2.50%
SURPRISED:	1.17%
HAPPY:	0.52%
ANGRY:	0.04%

LightGBM

CALM:	84.62%
SAD:	7.30%
ANGRY:	2.99%
NEUTRAL:	1.87%
SURPRISED:	1.77%
FEAR:	1.09%
HAPPY:	0.20%
DISGUSTED:	0.16%

Figure 3: User Interface #2 - Demonstration of Classification

The user interface component feeds into our pre-processing section. Pre-processing converts the input .mp4 video file into an array of image files that are later used for emotion classification. Figure 4 illustrates the several components that make up the preprocessing pipeline. Processing begins by stripping the audio data from the input video file. Once separated, silence detection clips silence from the beginning and the end of the file, this is important as classification information cannot be acquired from silent frames. The video clip is then cropped accordingly to ensure that video frames match up with respective audio frames.

The resulting visual video data is sampled at 0.6s intervals to produce a series of still facial images. Likewise, the audio data is also clipped into 0.6 second sections and Mel-Frequency Cepstral (MFC) analysis is performed to create a spectrogram image. The MFC analysis process is described in further detail in the Final Design Decisions section below.

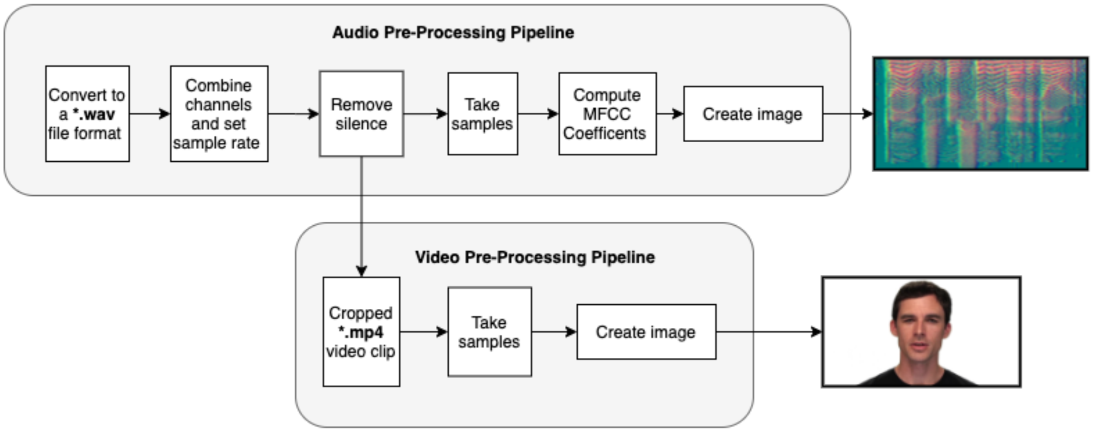

Audio Pre-Processing Pipeline
Convert to a *.wav file format
Combine channels and set sample rate
Remove silence
Take samples
Compute MFCC Coefficents
Create image
Video Pre-Processing Pipeline
Cropped *.mp4 video clip
Take samples
Create image

Figure 4: Pre-Processing Pipeline Overview

Once pre-processing is complete, the series of sampled facial images is sent to an Amazon Web Services (AWS) tool called Rekognition [7]. Recognition analyzes each image individually, and using an internal machine learning network, returns an array of emotional predictions per image. This data is saved for later use in our pipeline.

In parallel, the mel-spectral images produced in the pre-processing stage are fed into the audio neural network component of our system. The final version of the network runs the spectral images through pretrained Alexnet's feature extraction, and then through a custom 2 layer linear classifier. This outputs a tensor of the 8 predictions per emotion category for each image.

After AWS and our audio neural network have processed and classified the pre-processed images, we have a set of data with emotion predictions for each sample taken from the original mp4 video. The final stage in classification is to combine these results in a meaningful way, in order to produce one final classification. This final step is done using a gradient boosting decision tree. The tree, implemented using the LightGBM framework, creates a final stronger classification from the provided audio-based and image-based classification. The framework is trained on our data set with scripts that automatically attempt multiple plausible parameter combinations to produce the best classifier.

The final classification predictions are sent back to the user interface, where they are displayed. In the case of the educational interface, only the final top prediction is displayed. For the purposes of our classification pipeline demonstration interface, a significantly greater quantity of data is saved throughout the process of classification, and is displayed to the user.

One important component of our system that is not shown through these 4 modules is the backend pipeline that combines them all. This pipeline is a script that oversees the execution of all the components of the pipeline, as well as their outputs and inputs. Due to the design decisions that we made early on - that is, to create modular components - we were able to build a pipeline to put them all together seamlessly. From the user interface, the entire pipeline runs from one click to trigger it. This also allowed us to automatically save important intermediary information, such as emotion predictions from AWS and our audio neural network, in order to display in our demo user interface, and also for accuracy and quality checks. With all of the components of our system, and the pipeline that integrates them, we are able to achieve our final system design and product.

Final Design Decisions (Author: O. Roscoe)

For the user interface, we began with only one interface design - one for our target users, children with autism. Towards the end of our project development, we realized that we had an abundance of important data that our system was generating that we did not have the opportunity to display. Under the recommendation of our supervisor, we decided to create a second interface where we could display more of the technical data from our system. It was important to distinguish these two interfaces, as their target users are very different.

Initially, we had included data capture in the form of live audio and video as a part of our project scope. While building our system, we realized the limitations of such an input feed. These limitations include a lack of controlled environment, lack of emotional consistency, and lack of quality control. For the purpose of meeting our project goals, we prioritized the development of the classification components using mp4 video inputs, and so we narrowed our scope to take an input of this kind as well.

We decided to implement Mel-Frequency Cepstral analysis for audio feature extraction. MFC analysis is often used in tonal recognition to identify the patterns and components associated with human speech by modeling the natural filters inside the human ear [8]. This is particularly useful to the design as MFC can identify prosodic components associated with emotion and social cues in speech (such as tone, pitch, and power) by using the rate of change between MFCC [9][8]. Therefore, since some aspects of emotions are perceived largely through tone instead of specific words, we decided to use MFCs to process audio data [10].

The final audio network architecture was decided after extensive testing and analysis was done with other models such as simple CNNs. Through testing various architectures by adding layers, dropout, and different activations, using pretrained Alexnet feature extraction produced the highest accuracy for validation and testing.

The image classification leveraged a pre-existing solution, AWS Rekognition, to tag the images produced by pre-processing with emotions [7][11]. AWS was compared with a baseline AlexNet-based solution, but significantly outperformed it in a limited test so the team decided to continue with Rekognition.

The final classification used LightGBM, a gradient-boosted decision tree framework. This was done because it is able to rapidly train on large datasets and provide a strong prediction, avoiding manual attempts to guess at the relationships in the data. This was also able to significantly outperform a baseline model which involved taking the emotion with the highest score after averaging the image-based and audio-based prediction (71.25% final test accuracy vs. 24.134% accuracy).

Testing and Verification

Verification Matrix (Author: M. Muldoon)

Change?	ID	Requirement	Verification Result and Proof	Requirement Verification Method			
				Similarity	Review of Design	Analysis	Test
	1.1	Output correctly classified emotions with minimum 60% test accuracy, with an objective of 80% accuracy.	**Pass**: Final accuracy of 71.25% obtained on test set. See Appendix C, Figure C8 & C9				**X**
	1.2	Accept video and audio data as input.	**Pass**: See Appendix C, Figure C1, Figure C2 & Figure C3			**X**	
Modified to 60% to align with 1.4	1.3	Achieve minimum 60% accuracy in emotion classification from image classification.	**Pass**: Final accuracy of 62% obtained on test set. See Appendix C, Table C1				**X**
	1.4	Achieve minimum 60% accuracy in emotion classification from audio classification.	**Fail**: A final accuracy of 45% was achieved. See Appendix C, Figure C4				**X**
	2.1	Achieve a trained network for 6 human emotions (happy/joy, neutral/calm, sad, angry, fear,	**Pass**: Training was achieved for a total of 8 human emotions (happy, sad, calm, angry,				**X**

		disgust).	confused, disgusted, fear, surprised). See Appendix C, Figure C5.				
	2.2	Input will be classified within 10 seconds after being uploaded to the application [5].	**Pass**: Classification time is ~ 5 seconds. See Appendix C, Figure C6 and Figure C7.				**X**

Final Test Result (Author: S. Tijanic)

The test criteria that we used is a combination of qualitative and quantitative verification methods that target all components of our system. From the user interface, we met our objective of accepting video and audio data, as can be seen in Figures C1, C2 and C3 in Appendix C, where our system accepts and stores an mp4 file. The system itself, namely the backend classification pipeline that is triggered from the front end, was tested to meet our objective of a 10 second response window. We have found that an average response time for our classification is 5 seconds, using Python timing tools as can be seen in Figures C6 and C7 of Appendix C. This makes our final classification time half the length of our goal time, which is a positive achievement, as low response time is an asset in user applications.

In terms of our emotion classification, it was important for us to be able to distinguish between at least 6 different human emotions. After building our components and integrating them, we were able to work with 8 total human emotions - surpassing our initial requirement. This can be seen from the classification pipeline data in Figure C5 in Appendix C. In classifying these emotions, we were able to achieve approximately 62% accuracy in image classification using AWS Rekognition. Our final audio neural network classification accuracy for validation is 44.89%, this unfortunately doesn’t meet our 60% requirement. Combining the predictions from our audio and image classification, our gradient boosting tree produces a final prediction accuracy of 71.25%, which meets our minimum accuracy requirement for the system, but is below our 80% accuracy objective.

Summary and Conclusions

Conclusion (Author: S. Tijanic)

The F.A.C.E (Facial and Audio-based Classification of Emotion) application that we created is a combination of audio and video data processing, neural networks, decision trees, front end interfaces and backend pipelines. The project encompasses the challenges of machine learning, working with audio data, attempting to decipher human emotion and catering to a niche and important target user.

Our major goal of creating this application, and having it classify human emotion from video input, was achieved, as seen in our verification matrix. We believe that our verification methods were able to accurately depict the success of our project, as they encompassed all of the components, and both qualitative and quantitative measurements. We feel that our project was completed through all of the components that we intended to create, however we acknowledge that there is always room to improve. In particular, our audio classification component was a challenging area that we set out to solve. We know that both in our project, and in the field of research centered around this problem, there is a lot of opportunity to improve analysis methods and increase the accuracy of emotional classification. We see this room for improvement as a positive result of our project, and we believe that a good foundation of work and knowledge has been set to allow for further improvement. In the end, we were able to validate that it is possible to achieve good accuracy of classification of human emotion from video input, and this was our project’s primary goal.

From the onset of this project, our team agreed on an important design principle - modularity. With a system composed of many smaller components, and the need to integrate all of these components into one cohesive application, it was integral that we build all of our pieces in small, versatile chunks. As can be seen from our system overview diagram (Figure 1), we compartmentalized our modules from the beginning. This allowed us to split work among our team, prevent bottlenecks and blockers, and create modular components. When it came time to tie all of our components together into one pipeline, the process was intuitive and predictable because we had designed our system to fit together from the beginning.

Another key design decision that we made early on in our process was to use AWS Recognize for our facial image classification. We choose to use a pre-existing service for one of our components in order to allow us to focus more time and resources into other areas of the project. As we predicted, audio processing, analysis and classification was one of the most challenging components of our project. It was important that we scoped our work accordingly, allowing us to work on a challenging piece, and use existing solutions elsewhere.

The background behind this project came from the hands-on experience that our team has had in working with children with autism. This created a strong motivation for us to create a tool that could be of use to children who have autism and struggle to identify the emotions of those around them. We wanted to create an application that could be useful to children and educators in the autistic community in both it's functionality and useability. Our research-driven user interface aims to cater our technology to this group of individuals, while our emotion classification pipeline aims to provide a service that is lacking.

The way that we envision our application to be used is between educational assistants and children with autism. Educational assistants may guide children to upload videos, see emotion results, and learn through patterns, practice and observation. With research showing that positive results can be achieved through such educational means and tools [4], we believe that there is a place for an application like ours in the world of education.

Future work for this project consists of two primary categories - the usage of our tool, and the technical research and improvements in our system. First, it is imperative that we collect feedback from educational assistants and children with autism on the user interface of our application, and how it behaves. This step would allow us to make adjustments, and cater our product to the users that it is designed for. We have a limited ability to know what we are missing or how we can improve without talking to our users directly, so this is an essential future step.

Further, we know that research in the area of audio processing and machine learning is still very new. It would be important for our team to invest more time in developing our audio neural network, and consulting with others who have done similar work in order to learn more and share our findings. As both a way of achieving better prediction accuracy, and helping to advance this area of study, spending more time on our audio classification component is an important future step in this project. We believe that some of the issues in audio accuracy come from the way the video clips were sampled, since not every word when speaking accurately represents the emotion being portrayed.

To conclude, the team has learned a great deal from every component in this project, and from working with each other. We are happy to have been able to create a working application that achieves the goal that we set out for, and we are excited at the prospects of some of our work helping other people in education and academia.

References

[1] American Psychological Association, Autism. [Online] Accessed May 11, 2019. Available: https://www.apa.org/topics/autism/

[2] Pavlov, Nikolay. (2014). User Interface for People with Autism Spectrum Disorders. Journal of Software Engineering and Applications. 07. 128-134. 10.4236/jsea.2014.72014.

[3] Szilvia Papp (2006), "A Relevance-Theoretic Account of the Development and Deficits of Theory of Mind in Normally Developing Children and Individuals with Autism", *University of Portsmouth*. [Online]. Accessed May 11, 2019. Available: https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/09593543/v16i0002/141_araotddcaiwa.xml

[4] Michelle R. Kandalaft, Nyaz Didehbani, Daniel C. Krawczyk, Tandra T. Allen, Sandra B. Chapman (09 May 2012), "Virtual Reality Social Cognition Training for Young Adults with High-Functioning Autism", *Journal of Autism and Developmental Disorders*. [Online]. Accessed May 14 2019. Available: https://link.springer.com/article/10.1007/s10803-012-1544-6

[5] Baraković, S., & Skorin-Kapov, L. (2017). Modelling the relationship between design/performance factors and perceptual features contributing to quality of experience for mobile web browsing. *Computers in Human Behavior, 74*(Complete), 311-329. Accessed September 30, 2019. Available: https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/07475632/v74icomplete/311_mtrbdfoefmwb.xml

[6] Ke, G., , Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. , Lui, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree *Neural Information Processing Systems.* Accessed: November 29, 2019. Available: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf
[7] Amazon Rekognition (2019), Amazon Rekognition: Easily add intelligent image and video analysis to your applications. [Online] Accessed May 24, 2019. Available: https://aws.amazon.com/rekognition/

[8] K. S. Prahallad. Speech Technology: A Practical Introduction. Class Lecture, Topic: "Spectrogram, Cepstrum and Mel-Frequency Analysis". Carnegie Mellon University, International Institute of Information Technology, Hyderabad. Hyderabad, India. July 26, 2011.

[Online]. Accessed May 24, 2019. Available:
https://sites.google.com/site/kishoreprahallad/presentations

[9] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir (Mar. 2019), “Audio-visual emotion fusion (AVEF): A deep efficient weighted approach,” *Information Fusion*, vol. 46, pp. 184–192. [Online]. Available:
https://www-sciencedirect-com.myaccess.library.utoronto.ca/science/article/pii/S1566253518300733?via%3Dihub

[10] E. Liebenthal, D. Silbersweig and E. Stern, "The Language, Tone and Prosody of Emotions: Neural Substrates and Dynamics of Spoken-Word Emotion Perception", *Frontiers in Neuroscience*, vol. 10, 2016. Available: 10.3389/fnins.2016.00506.

[11] Microsoft Azure (2019), Face. [Online] Accessed May 25 2019. Available:
https://azure.microsoft.com/en-us/services/cognitive-services/face/

[12]"Amazon Rekognition – Pricing - AWS", Amazon Web Services, Inc., 2019. [Online]. Available: https://aws.amazon.com/rekognition/pricing/?nc=sn&loc=4. [Accessed: 10- Oct- 2019]

[13] "Amazon Elastic Inference Pricing - Amazon Web Services", Amazon Web Services, Inc., 2019. [Online]. Available: https://aws.amazon.com/machine-learning/elastic-inference/pricing/. [Accessed: 10- Oct- 2019]

Appendices

Appendix A: Gantt Chart History

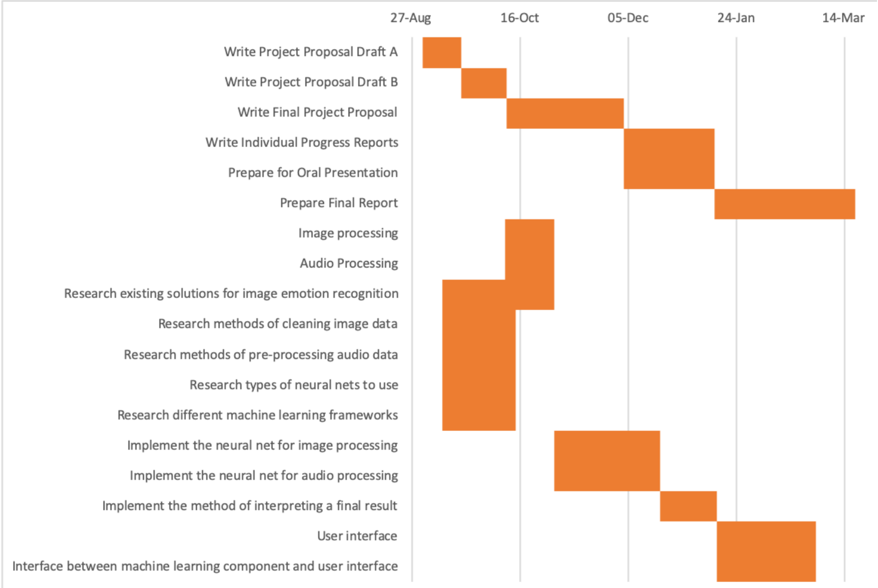


Figure A1: Project Proposal - Gantt Chart

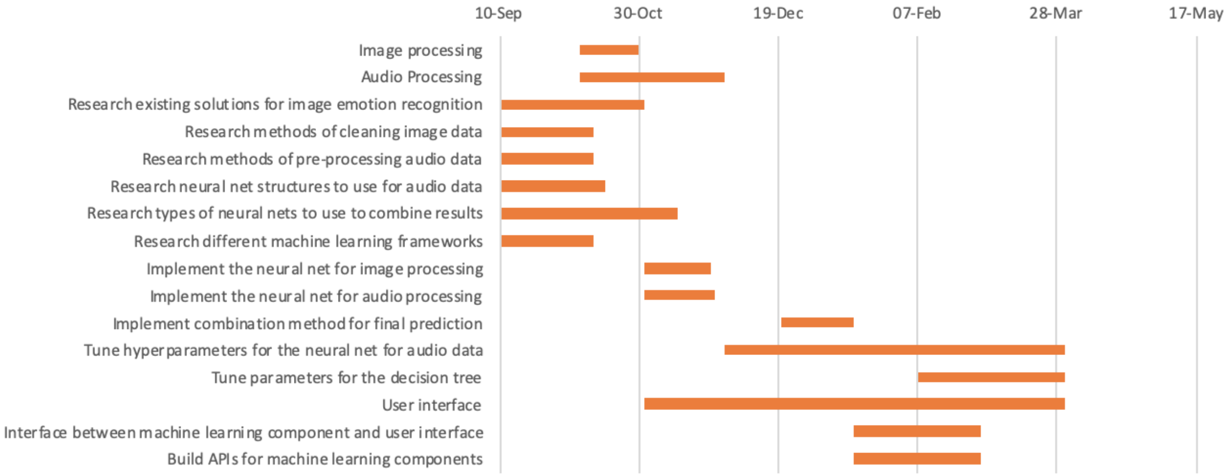


Figure A2: Final Report - Gantt Chart

Appendix B: Financial Plan

The financial expenditures on the project can be divided into two main categories, the hardware required for the project and the cloud computing/computational resources required. Between these two categories, cloud computing spending is expected to heavily dominate the expected expenses. This is because it may be needed to help properly label datasets and will be needed for training the machine learning solution that forms the central part of this project.

We ran our datasets Amazon Web Services (AWS) Rekognition. Our data consists of approximately 3 data sets containing 7000 entries of 3s clips of audio/video. Assuming that we run this entire dataset through AWS speech/video services and cache (i.e. save) those results to minimize cost, our expenditure would be $105 for visual data through AWS Rekognition. Additionally, we needed to train our neural network. This is a fairly intensive process that is greatly accelerated by advanced computational hardware that is available from cloud computing providers. Following a pricing structure provided by AWS through their Elastic Inference acceleration service, one hour of computation time on a powerful, accelerated computing instance costs approximately $0.215. We expect that over the course of the project our total usage will be approximately 750 hours of computation time, approximately one month. This results in a total computation cost of $161.25. Thus, in total we expect to spend $266.25 on cloud services.

Additionally, there are the hardware costs for the project. All team members have personal computers that they will be able to use for development purposes, which have web cameras capable of recording video.

Upon the completion of this project, we have been able to keep our final cost for AWS Rekognition and AWS Elastic Interface at $0. For the AWS Recognition service, each member of the team was given a free usage quota upon creating an AWS account, and we were able to run all of our data through this service within our free quotas. Regarding the AWS Elastic Interface, our project supervisor was able to provide us with access to accelerated computing resources in order to train our neural networks. This allowed us to complete the components of our project requiring greater computational power without the use of AWS Elastic Interface.

Table B1 - Expected Financial Expenditure for Project

Item	Expected Number of Units	Cost per Unit	Expected Cost	Actual Cost
AWS	6300 (3 data	$0.01667 per	$105	$0

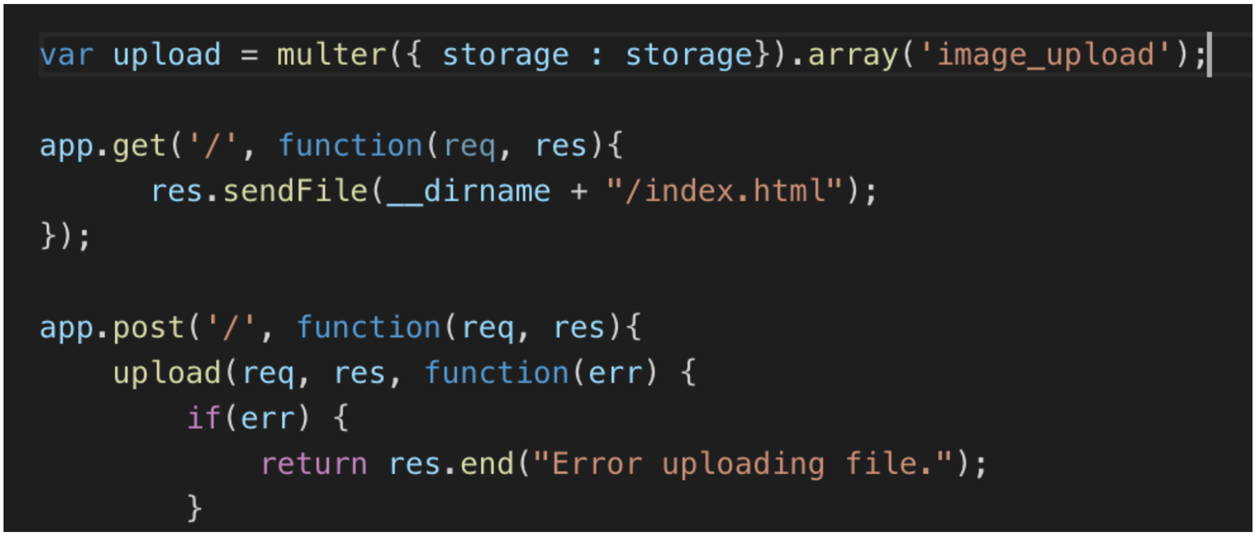
Rekognition [15]	sets x 7000 clips x 3s/clip)	minute		
AWS Elastic Interface [17]	750 hours	$0.215	$161.25	$0
Student Time	2080 hours (4 students x 10 hours/week x 52 weeks)	$30	$62 040	$0
Total	N/A	N/A	$62,306.25	$0

Appendix C: Validation and Acceptance Tests

Validation and Acceptance Tests from Project Proposal:
The project goal, as previously stated, is to develop a web application that can identify emotions and social cues. As such, the success of this project can be measured by the ability to execute this functionality, as well as the various levels of accuracy, detail and presentation of doing so.

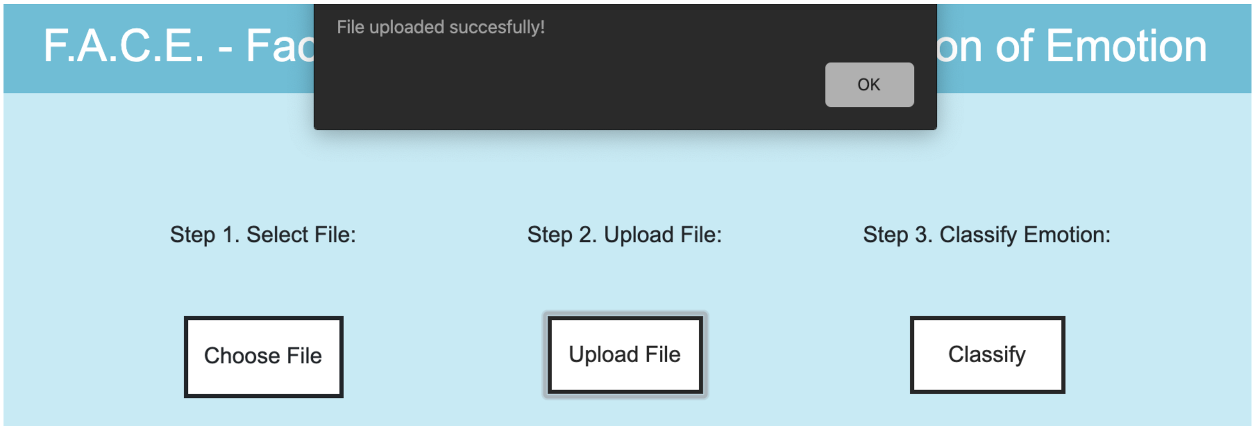
Table C1: Verification Tests on Project Requirements and Objectives

ID	Project Requirement	Verification Method
1.1	Output correctly classified emotions from a pre-defined group of human emotions.	TEST: Predetermined labels of validation data will be compared against the solution output to determine whether identification of human emotion is correct.
1.2	Accept video and audio data as input.	REVIEW of DESIGN: Input parameters will be verified as video and audio.
ID	Project Objectives	Verification Method
2.1	Correctly classify between 3-5 human emotions.	TEST: Validation data will produce accurate classification for a minimum of 3 human emotions.
2.2	Minimize time required to classify input.	TEST: Average application processing time will be calculated and compared against 10 seconds [5].



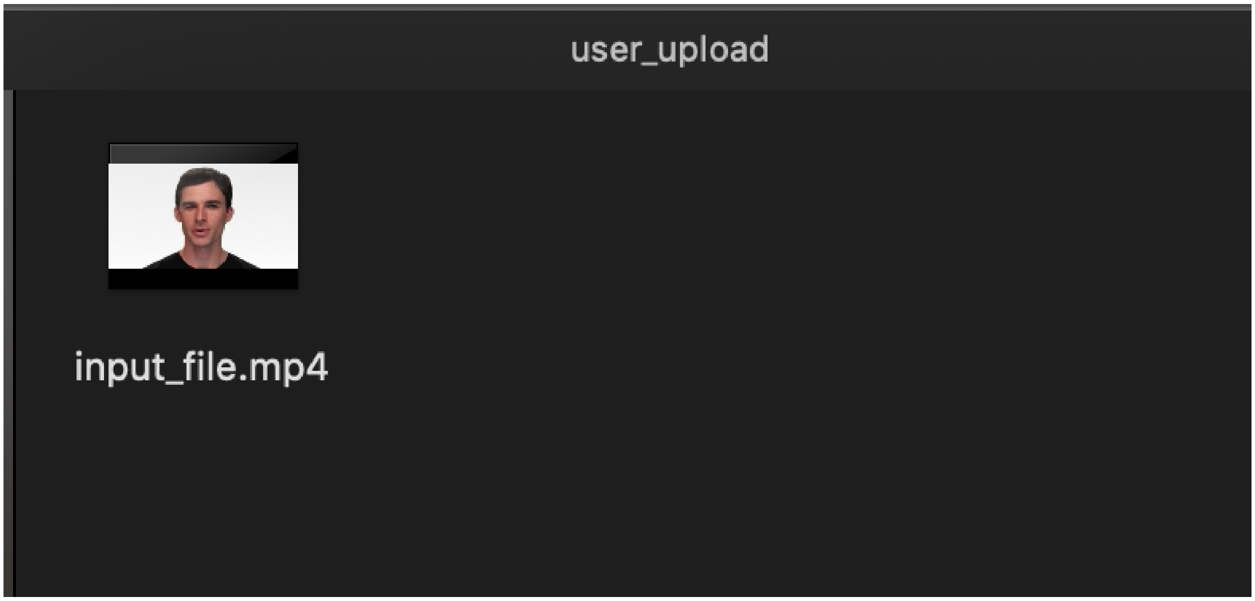
```
var upload = multer({ storage : storage}).array('image_upload');

app.get('/', function(req, res){
      res.sendFile(__dirname + "/index.html");
});

app.post('/', function(req, res){
    upload(req, res, function(err) {
        if(err) {
            return res.end("Error uploading file.");
        }
```

Figure C1 - File Upload Code Snippet



File uploaded succesfully!

Figure C2 - File Uploaded onto Front-end UI


user_upload

input_file.mp4

Figure C3 - File Accepted from Front-end to Back-end

Table C1 - Evaluations on Sample Dataset

Evaluated Dataset/Approach	Accuracy (%)	Number of Samples
Total Data Set (Actual Emotion = Top Prediction by AWS Rekognition)	44.32	5869
Total Data Set (Actual Emotion in Top 2 Predictions by AWS Rekognition)	64.08	5869
LightGBM Training Set Accuracy (Used to train the LightGBM implementation)	93.69	3534
LightGBM Validation Set Accuracy (Used for early stopping and validation)	59.04	625
LightGBM Test Set Accuracy (Completely unseen data)	61.98	626

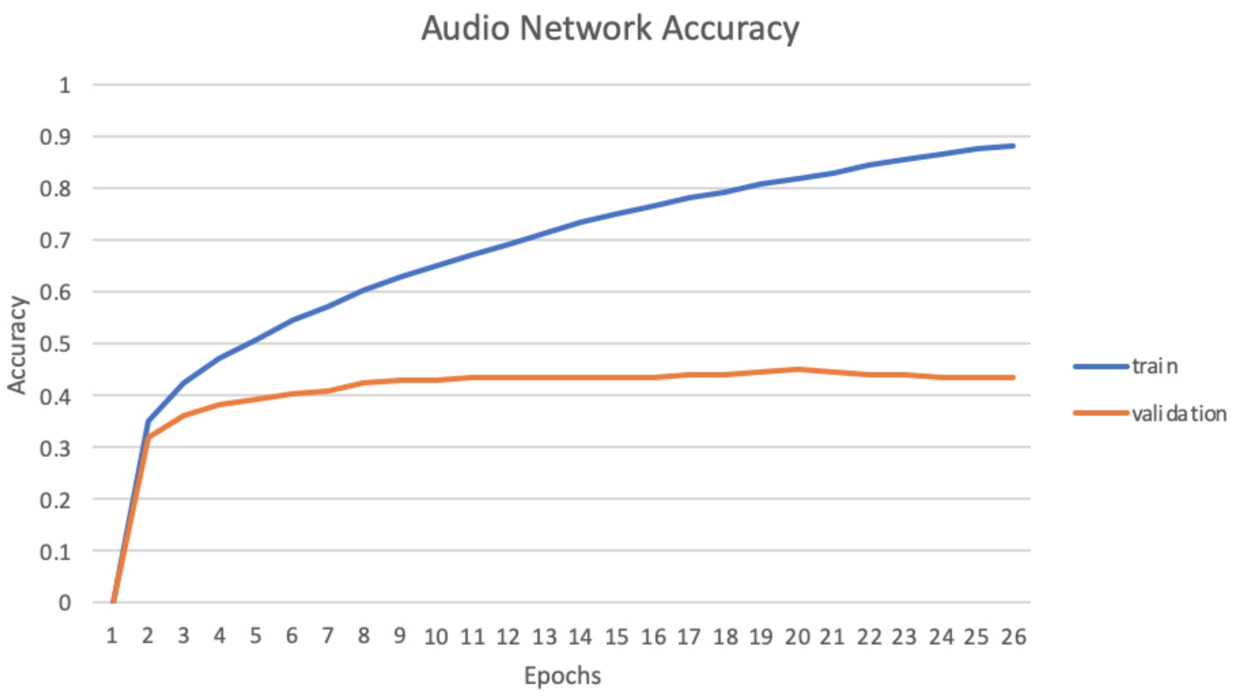
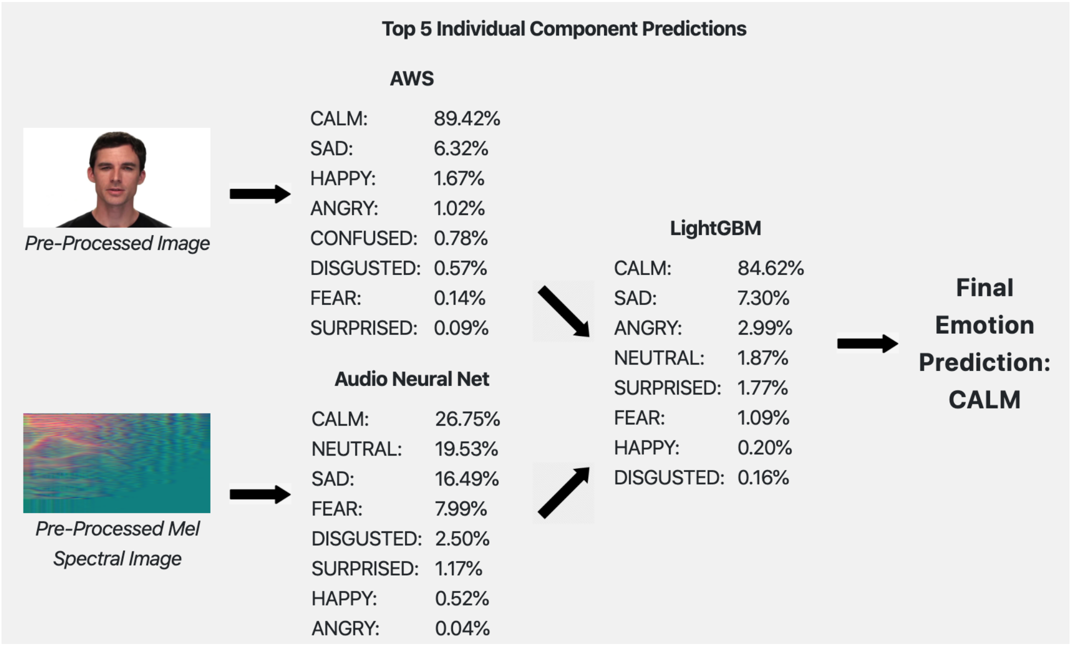


Figure C4 - Training Accuracy vs Epochs for Audio Neural Net



Top 5 Individual Component Predictions

AWS

CALM:	89.42%
SAD:	6.32%
HAPPY:	1.67%
ANGRY:	1.02%
CONFUSED:	0.78%
DISGUSTED:	0.57%
FEAR:	0.14%
SURPRISED:	0.09%

Audio Neural Net

CALM:	26.75%
NEUTRAL:	19.53%
SAD:	16.49%
FEAR:	7.99%
DISGUSTED:	2.50%
SURPRISED:	1.17%
HAPPY:	0.52%
ANGRY:	0.04%

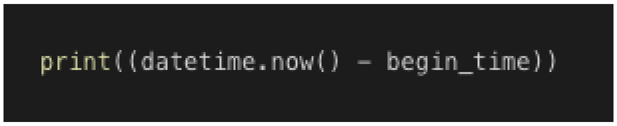
LightGBM

CALM:	84.62%
SAD:	7.30%
ANGRY:	2.99%
NEUTRAL:	1.87%
SURPRISED:	1.77%
FEAR:	1.09%
HAPPY:	0.20%
DISGUSTED:	0.16%

Final Emotion Prediction: CALM

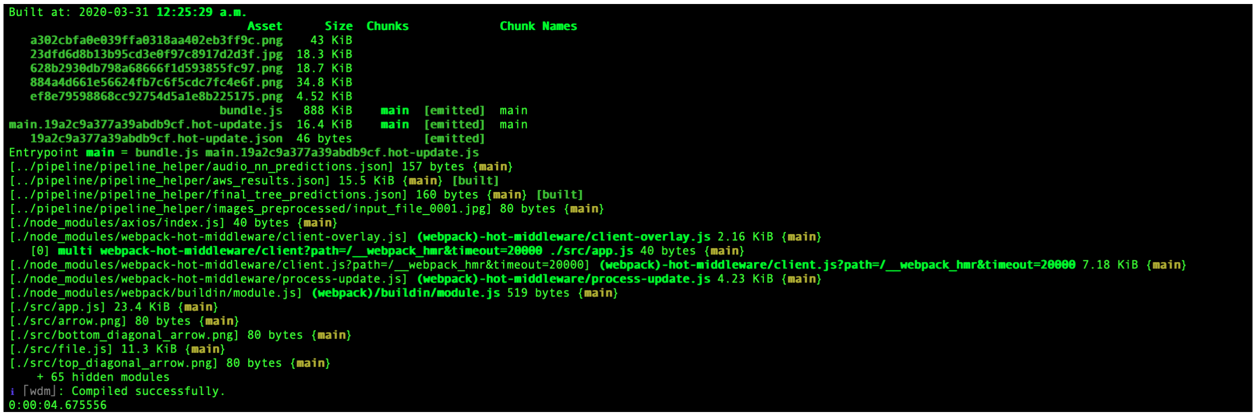
Figure C5 - Training Accuracy vs Epochs for Audio Neural Net

Figure C5 shows the classification pipeline data. It is shown that a total of 8 human emotion categories are used for classification throughout the system. AWS returns 8 emotion categories, our audio neural network classifies 8 emotion categories, and the gradient boosting tree also shows 8 final emotion classifications. The top result becomes the final prediction.



```python
print((datetime.now() - begin_time))
```

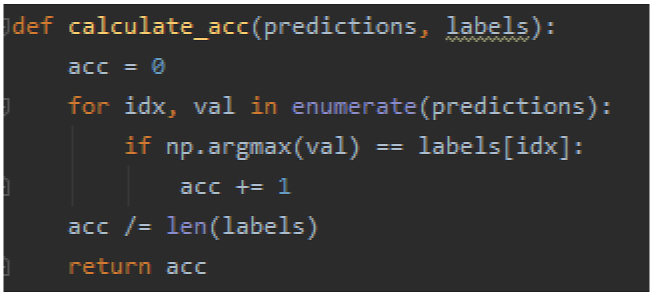
Figure C6 - Python Code Used to Time System



```
Built at: 2020-03-31 12:25:29 a.m.
                                   Asset      Size  Chunks             Chunk Names
   a302cbfa0e039ffa0318aa402eb3ff9c.png    43 KiB
   23dfd6d8b13b95cd3e0f97c8917d2d3f.jpg  18.3 KiB
   628b2930db798a68666f1d593855fc97.png  18.7 KiB
   884a4d661e56624fb7c6f5cdc7fc4e6f.png  34.8 KiB
   ef8e79598868cc92754d5a1e8b225175.png  4.52 KiB
                               bundle.js   888 KiB    main  [emitted]  main
main.19a2c9a377a39abdb9cf.hot-update.js  16.4 KiB    main  [emitted]  main
   19a2c9a377a39abdb9cf.hot-update.json  46 bytes          [emitted]
Entrypoint main = bundle.js main.19a2c9a377a39abdb9cf.hot-update.js
[../pipeline/pipeline_helper/audio_nn_predictions.json] 157 bytes {main}
[../pipeline/pipeline_helper/aws_results.json] 15.5 KiB {main} [built]
[../pipeline/pipeline_helper/final_tree_predictions.json] 160 bytes {main} [built]
[../pipeline/pipeline_helper/images_preprocessed/input_file_0001.jpg] 80 bytes {main}
[./node_modules/axios/index.js] 40 bytes {main}
[./node_modules/webpack-hot-middleware/client-overlay.js] (webpack)-hot-middleware/client-overlay.js 2.16 KiB {main}
   [0] multi webpack-hot-middleware/client?path=/__webpack_hmr&timeout=20000 ./src/app.js 40 bytes {main}
[./node_modules/webpack-hot-middleware/client.js?path=/__webpack_hmr&timeout=20000] (webpack)-hot-middleware/client.js?path=/__webpack_hmr&timeout=20000 7.18 KiB {main}
[./node_modules/webpack-hot-middleware/process-update.js] (webpack)-hot-middleware/process-update.js 4.23 KiB {main}
[./node_modules/webpack/buildin/module.js] (webpack)/buildin/module.js 519 bytes {main}
[./src/app.js] 23.4 KiB {main}
[./src/arrow.png] 80 bytes {main}
[./src/bottom_diagonal_arrow.png] 80 bytes {main}
[./src/file.js] 11.3 KiB {main}
[./src/top_diagonal_arrow.png] 80 bytes {main}
    + 65 hidden modules
i ｢wdm｣: Compiled successfully.
0:00:04.675556
```

Figure C7 - Terminal Output Showing Execution Time

Figure C5 and C6 show the use of Python's datetime module to calculate the total execution time. The variable begin_time is set at the beginning of the classification pipeline, and the elapsed time is calculated and printed at the end, shown in Figure C5. Figure C6 shows that the total execution time is 4.68 seconds.



```python
def calculate_acc(predictions, labels):
    acc = 0
    for idx, val in enumerate(predictions):
        if np.argmax(val) == labels[idx]:
            acc += 1
    acc /= len(labels)
    return acc
```

Figure C8 - Code used to Calculate Accuracy of LightGBM Model

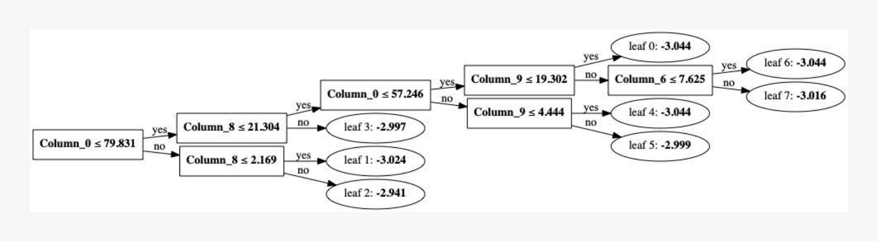
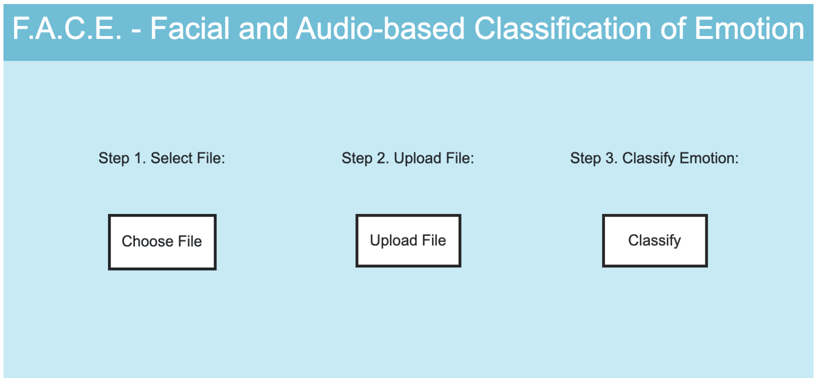


Figure C9 - Final LightGBM Model Diagraph



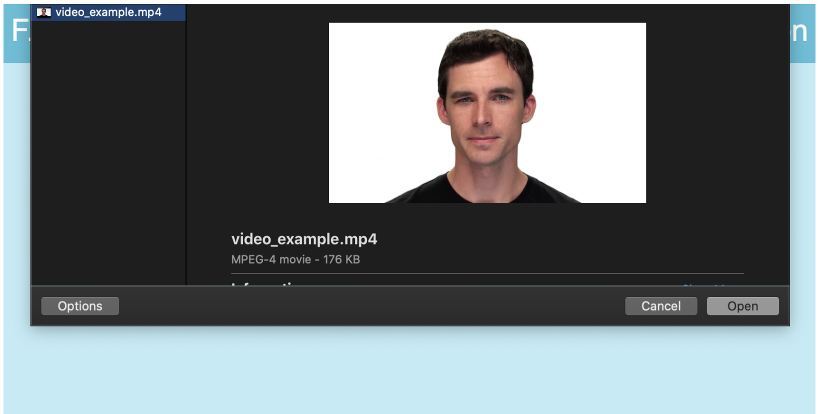
F.A.C.E. - Facial and Audio-based Classification of Emotion

Step 1. Select File:

Step 2. Upload File:

Step 3. Classify Emotion:

Figure C10: Home Screen of User Interface

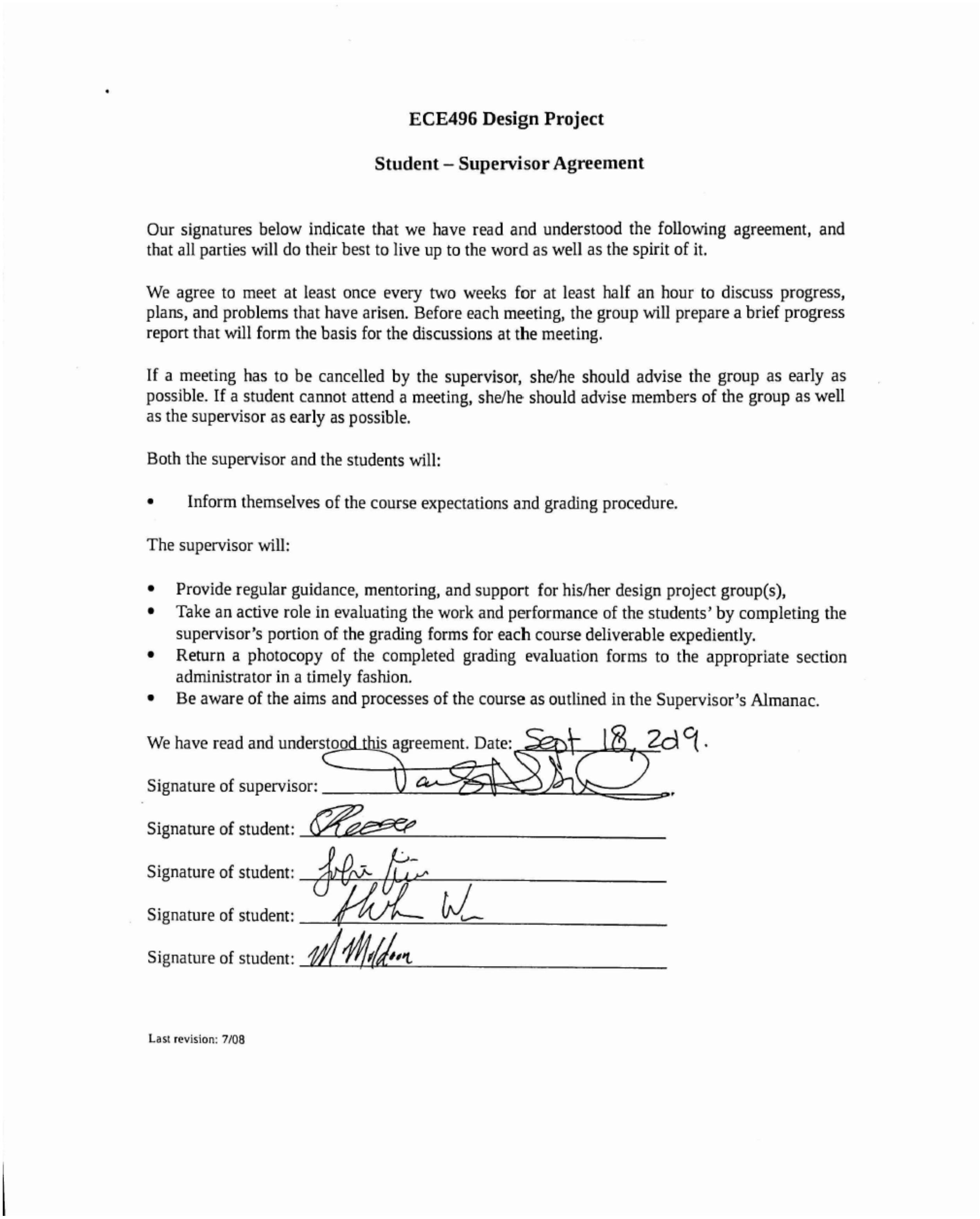


video_example.mp4

MPEG-4 movie - 176 KB

Figure C11: File Upload

Appendix D: Student-Supervisor Agreement Form



ECE496 Design Project

Student – Supervisor Agreement

Our signatures below indicate that we have read and understood the following agreement, and that all parties will do their best to live up to the word as well as the spirit of it.

We agree to meet at least once every two weeks for at least half an hour to discuss progress, plans, and problems that have arisen. Before each meeting, the group will prepare a brief progress report that will form the basis for the discussions at the meeting.

If a meeting has to be cancelled by the supervisor, she/he should advise the group as early as possible. If a student cannot attend a meeting, she/he should advise members of the group as well as the supervisor as early as possible.

Both the supervisor and the students will:

- Inform themselves of the course expectations and grading procedure.

The supervisor will:

- Provide regular guidance, mentoring, and support for his/her design project group(s),
- Take an active role in evaluating the work and performance of the students' by completing the supervisor's portion of the grading forms for each course deliverable expediently.
- Return a photocopy of the completed grading evaluation forms to the appropriate section administrator in a timely fashion.
- Be aware of the aims and processes of the course as outlined in the Supervisor's Almanac.

We have read and understood this agreement. Date: Sept 18, 2019.

Signature of supervisor: ______________________

Signature of student: ______________________

Signature of student: ______________________

Signature of student: ______________________

Signature of student: M Muldoon

Last revision: 7/08

Appendix E: Pre-Processing Overview

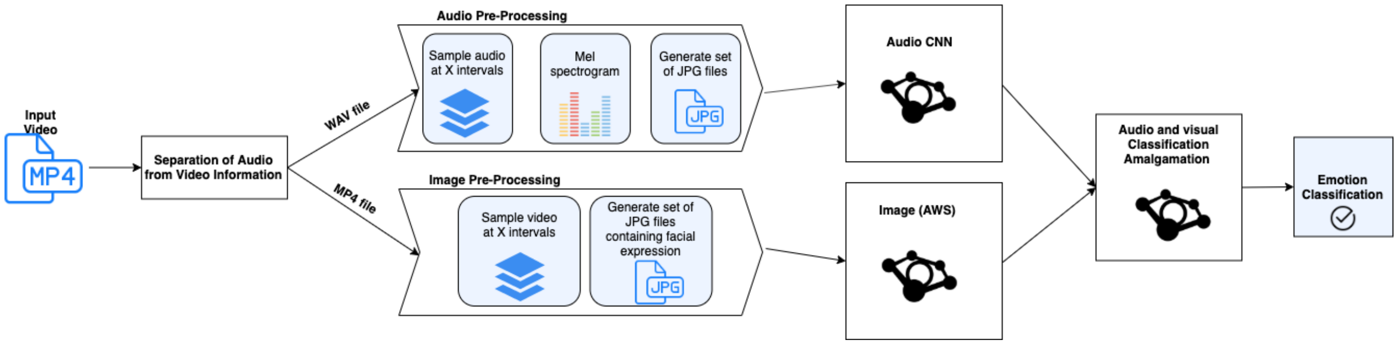
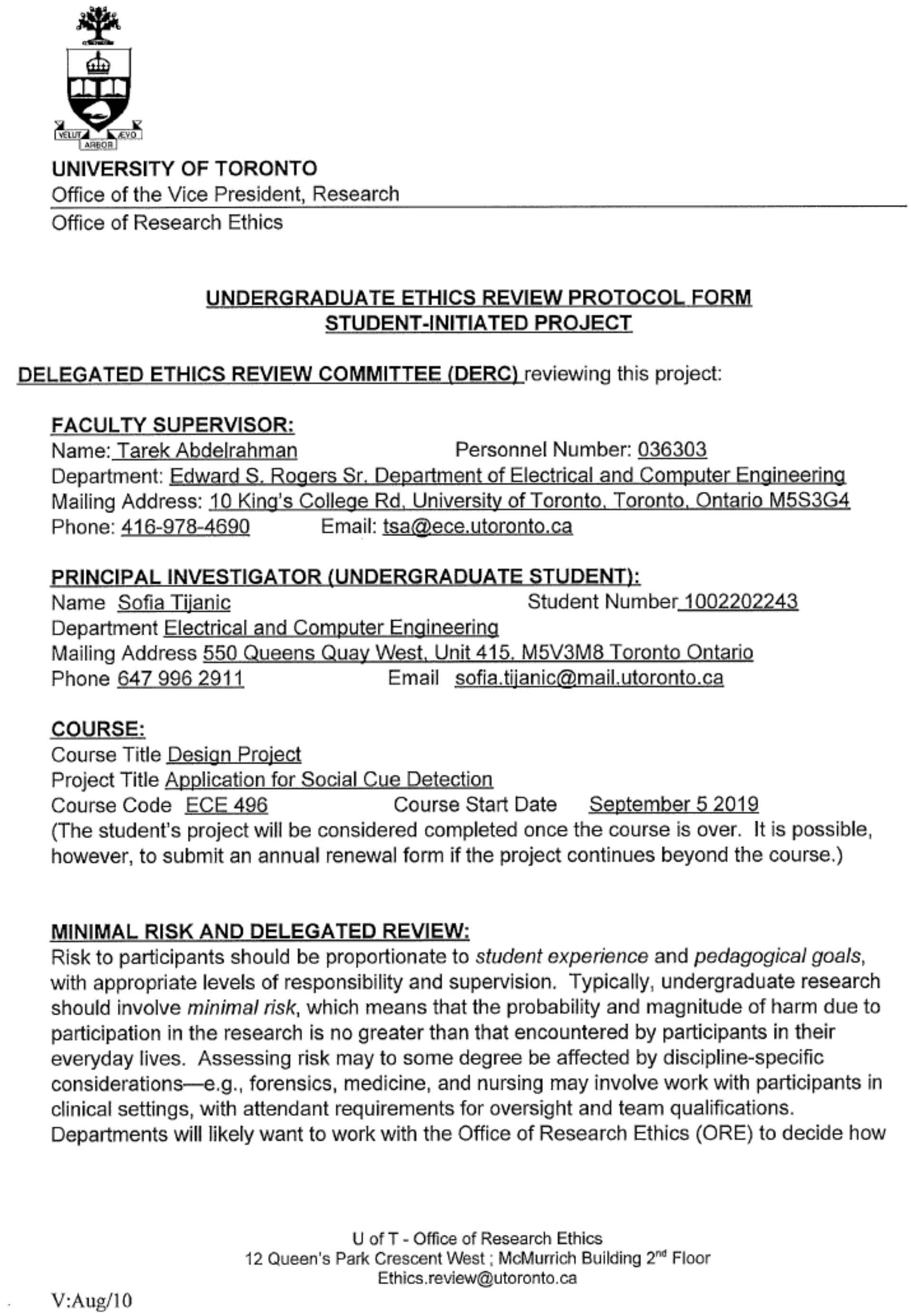

Input Video
MP4
Separation of Audio from Video Information
WAV file
MP4 file
Audio Pre-Processing
Sample audio at X intervals
Mel spectrogram
Generate set of JPG files
JPG
Image Pre-Processing
Sample video at X intervals
Generate set of JPG files containing facial expression
JPG
Audio CNN
Image (AWS)
Audio and visual Classification Amalgamation
Emotion Classification

Figure E1: Pre-Processing Overview

Schematic diagram containing the modules that the processing block is composed of.

Appendix F: Ethics Review Form



UNIVERSITY OF TORONTO
Office of the Vice President, Research
Office of Research Ethics

UNDERGRADUATE ETHICS REVIEW PROTOCOL FORM
STUDENT-INITIATED PROJECT

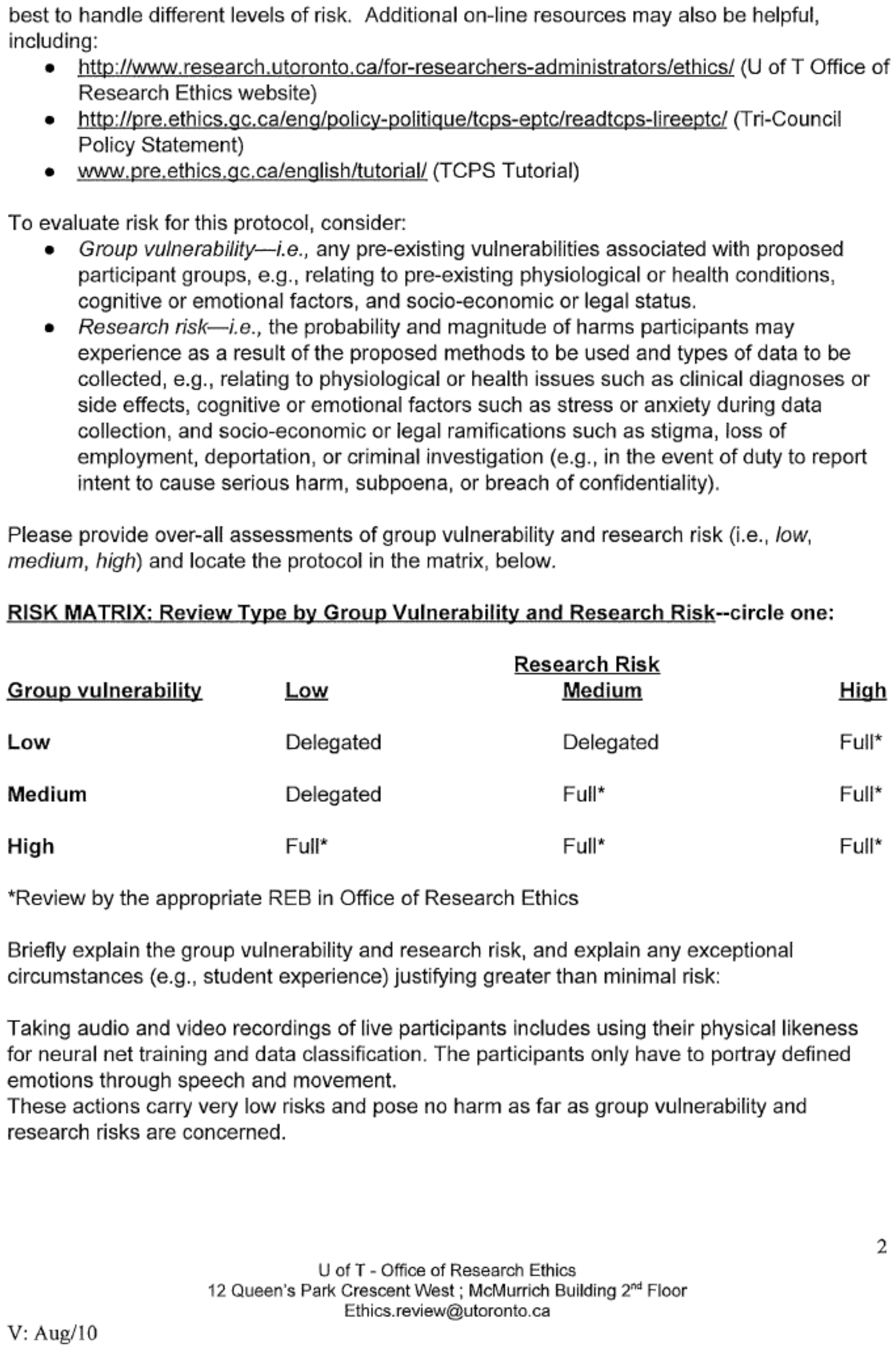
DELEGATED ETHICS REVIEW COMMITTEE (DERC) reviewing this project:

FACULTY SUPERVISOR:
Name: Tarek Abdelrahman Personnel Number: 036303
Department: Edward S. Rogers Sr. Department of Electrical and Computer Engineering
Mailing Address: 10 King's College Rd, University of Toronto, Toronto, Ontario M5S3G4
Phone: 416-978-4690 Email: tsa@ece.utoronto.ca

PRINCIPAL INVESTIGATOR (UNDERGRADUATE STUDENT):
Name Sofia Tijanic Student Number 1002202243
Department Electrical and Computer Engineering
Mailing Address 550 Queens Quay West, Unit 415, M5V3M8 Toronto Ontario
Phone 647 996 2911 Email sofia.tijanic@mail.utoronto.ca

COURSE:
Course Title Design Project
Project Title Application for Social Cue Detection
Course Code ECE 496 Course Start Date September 5 2019
(The student's project will be considered completed once the course is over. It is possible, however, to submit an annual renewal form if the project continues beyond the course.)

MINIMAL RISK AND DELEGATED REVIEW:
Risk to participants should be proportionate to *student experience* and *pedagogical goals*, with appropriate levels of responsibility and supervision. Typically, undergraduate research should involve *minimal risk*, which means that the probability and magnitude of harm due to participation in the research is no greater than that encountered by participants in their everyday lives. Assessing risk may to some degree be affected by discipline-specific considerations—e.g., forensics, medicine, and nursing may involve work with participants in clinical settings, with attendant requirements for oversight and team qualifications. Departments will likely want to work with the Office of Research Ethics (ORE) to decide how



best to handle different levels of risk. Additional on-line resources may also be helpful, including:

- http://www.research.utoronto.ca/for-researchers-administrators/ethics/ (U of T Office of Research Ethics website)
- http://pre.ethics.gc.ca/eng/policy-politique/tcps-eptc/readtcps-lireeptc/ (Tri-Council Policy Statement)
- www.pre.ethics.gc.ca/english/tutorial/ (TCPS Tutorial)

To evaluate risk for this protocol, consider:

- *Group vulnerability—i.e.,* any pre-existing vulnerabilities associated with proposed participant groups, e.g., relating to pre-existing physiological or health conditions, cognitive or emotional factors, and socio-economic or legal status.
- *Research risk—i.e.,* the probability and magnitude of harms participants may experience as a result of the proposed methods to be used and types of data to be collected, e.g., relating to physiological or health issues such as clinical diagnoses or side effects, cognitive or emotional factors such as stress or anxiety during data collection, and socio-economic or legal ramifications such as stigma, loss of employment, deportation, or criminal investigation (e.g., in the event of duty to report intent to cause serious harm, subpoena, or breach of confidentiality).

Please provide over-all assessments of group vulnerability and research risk (i.e., *low*, *medium*, *high*) and locate the protocol in the matrix, below.

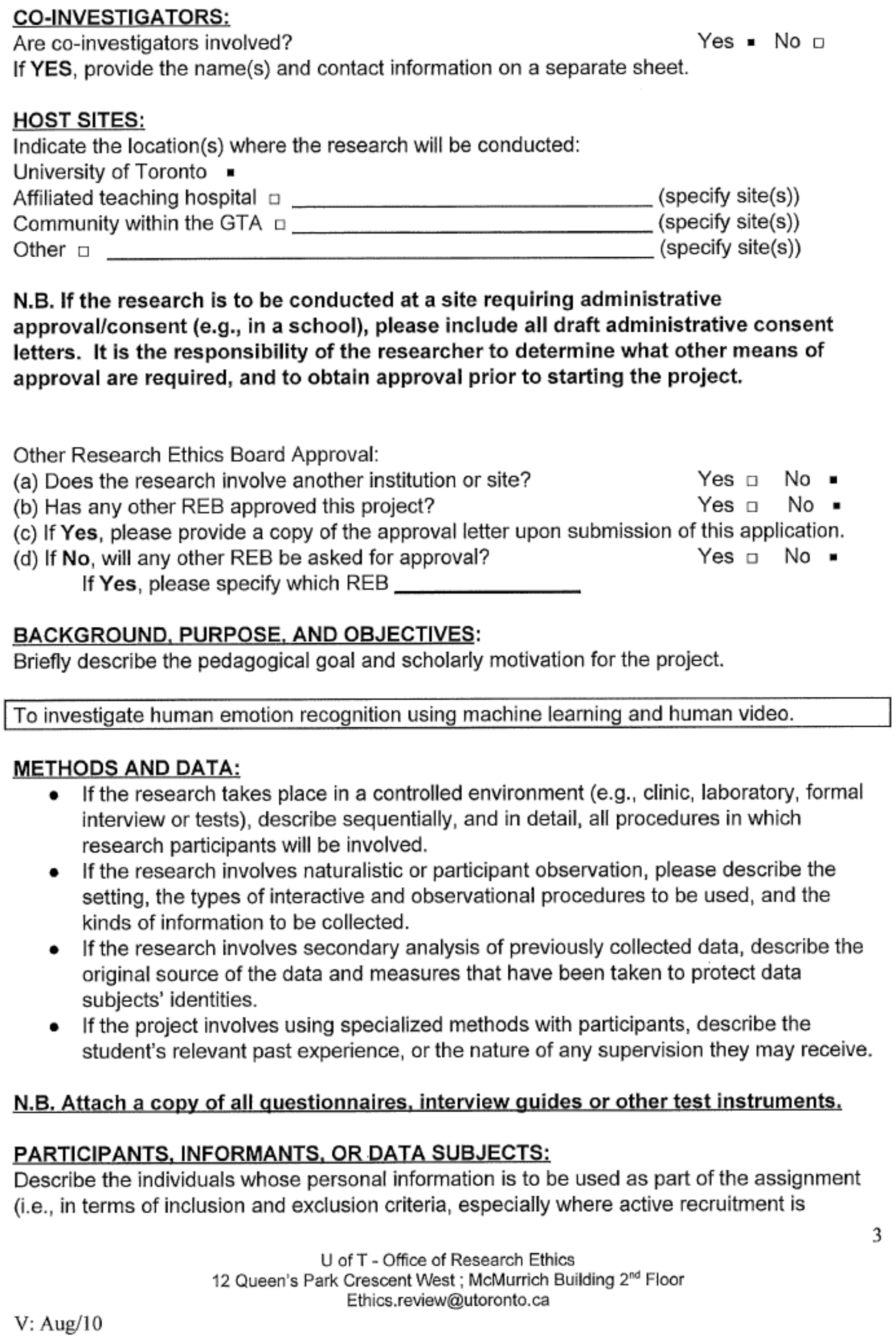
RISK MATRIX: Review Type by Group Vulnerability and Research Risk--circle one:

		Research Risk	
Group vulnerability	Low	Medium	High
Low	Delegated	Delegated	Full*
Medium	Delegated	Full*	Full*
High	Full*	Full*	Full*

*Review by the appropriate REB in Office of Research Ethics

Briefly explain the group vulnerability and research risk, and explain any exceptional circumstances (e.g., student experience) justifying greater than minimal risk:

Taking audio and video recordings of live participants includes using their physical likeness for neural net training and data classification. The participants only have to portray defined emotions through speech and movement.
These actions carry very low risks and pose no harm as far as group vulnerability and research risks are concerned.



CO-INVESTIGATORS:

Are co-investigators involved? Yes ■ No □

If **YES**, provide the name(s) and contact information on a separate sheet.

HOST SITES:

Indicate the location(s) where the research will be conducted:

University of Toronto ■

Affiliated teaching hospital □ ______________________________ (specify site(s))

Community within the GTA □ ______________________________ (specify site(s))

Other □ ____ (specify site(s))

N.B. If the research is to be conducted at a site requiring administrative approval/consent (e.g., in a school), please include all draft administrative consent letters. It is the responsibility of the researcher to determine what other means of approval are required, and to obtain approval prior to starting the project.

Other Research Ethics Board Approval:

(a) Does the research involve another institution or site? Yes □ No ■

(b) Has any other REB approved this project? Yes □ No ■

(c) If **Yes**, please provide a copy of the approval letter upon submission of this application.

(d) If **No**, will any other REB be asked for approval? Yes □ No ■

If **Yes**, please specify which REB ______________

BACKGROUND, PURPOSE, AND OBJECTIVES:

Briefly describe the pedagogical goal and scholarly motivation for the project.

To investigate human emotion recognition using machine learning and human video.

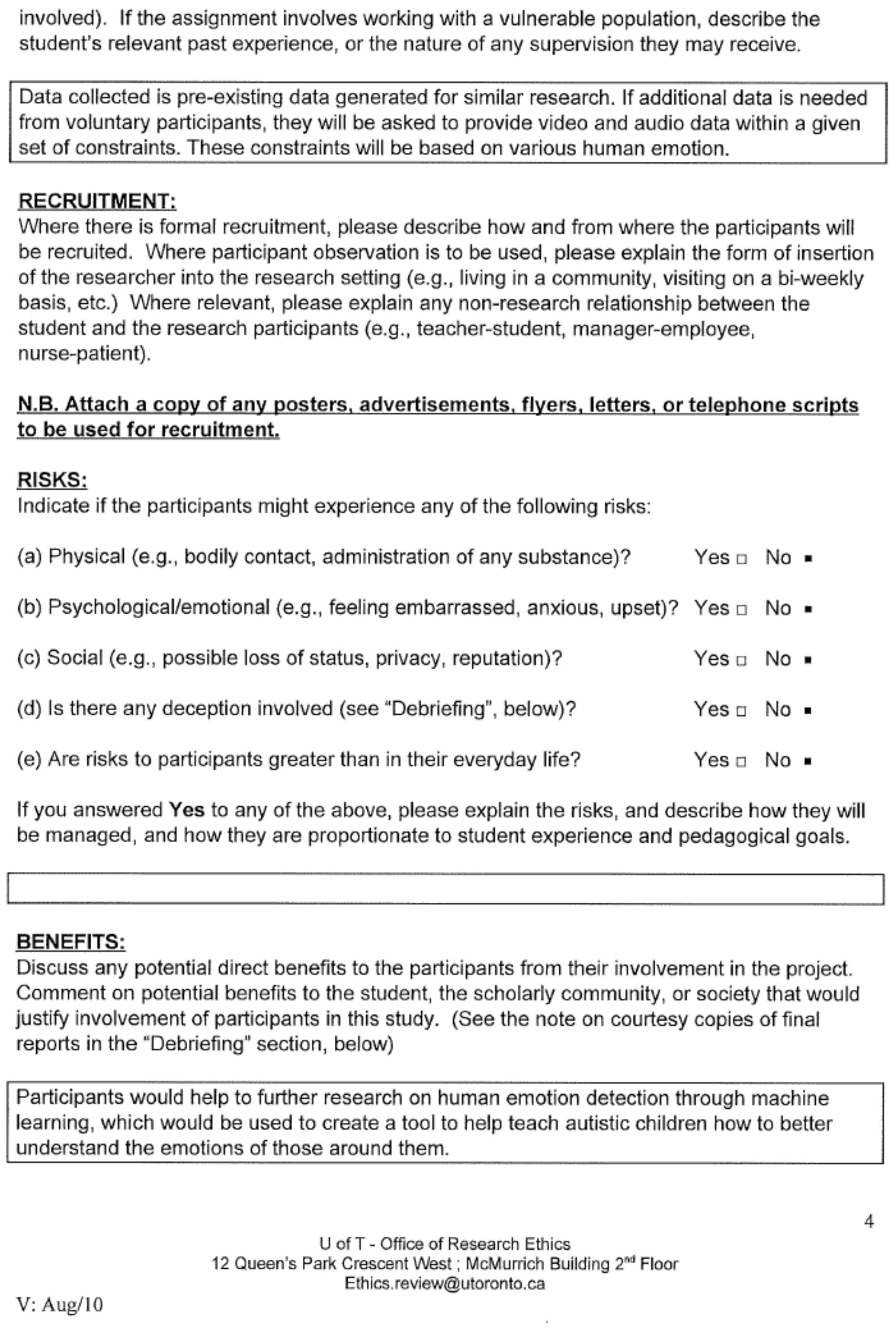
METHODS AND DATA:

- If the research takes place in a controlled environment (e.g., clinic, laboratory, formal interview or tests), describe sequentially, and in detail, all procedures in which research participants will be involved.
- If the research involves naturalistic or participant observation, please describe the setting, the types of interactive and observational procedures to be used, and the kinds of information to be collected.
- If the research involves secondary analysis of previously collected data, describe the original source of the data and measures that have been taken to protect data subjects' identities.
- If the project involves using specialized methods with participants, describe the student's relevant past experience, or the nature of any supervision they may receive.

N.B. Attach a copy of all questionnaires, interview guides or other test instruments.

PARTICIPANTS, INFORMANTS, OR DATA SUBJECTS:

Describe the individuals whose personal information is to be used as part of the assignment (i.e., in terms of inclusion and exclusion criteria, especially where active recruitment is



involved). If the assignment involves working with a vulnerable population, describe the student's relevant past experience, or the nature of any supervision they may receive.

Data collected is pre-existing data generated for similar research. If additional data is needed from voluntary participants, they will be asked to provide video and audio data within a given set of constraints. These constraints will be based on various human emotion.

RECRUITMENT:

Where there is formal recruitment, please describe how and from where the participants will be recruited. Where participant observation is to be used, please explain the form of insertion of the researcher into the research setting (e.g., living in a community, visiting on a bi-weekly basis, etc.) Where relevant, please explain any non-research relationship between the student and the research participants (e.g., teacher-student, manager-employee, nurse-patient).

N.B. Attach a copy of any posters, advertisements, flyers, letters, or telephone scripts to be used for recruitment.

RISKS:

Indicate if the participants might experience any of the following risks:

(a) Physical (e.g., bodily contact, administration of any substance)? Yes □ No ■

(b) Psychological/emotional (e.g., feeling embarrassed, anxious, upset)? Yes □ No ■

(c) Social (e.g., possible loss of status, privacy, reputation)? Yes □ No ■

(d) Is there any deception involved (see "Debriefing", below)? Yes □ No ■

(e) Are risks to participants greater than in their everyday life? Yes □ No ■

If you answered **Yes** to any of the above, please explain the risks, and describe how they will be managed, and how they are proportionate to student experience and pedagogical goals.

BENEFITS:

Discuss any potential direct benefits to the participants from their involvement in the project. Comment on potential benefits to the student, the scholarly community, or society that would justify involvement of participants in this study. (See the note on courtesy copies of final reports in the "Debriefing" section, below)

Participants would help to further research on human emotion detection through machine learning, which would be used to create a tool to help teach autistic children how to better understand the emotions of those around them.

COMPENSATION:

Will participants receive compensation for participation?		Yes □ No ■
	Financial	Yes □ No ■
	In-kind	Yes □ No ■
	Other	Yes □ No ■

(b) **If Yes**, please provide details.

(c) Where there is a withdrawal clause in the research procedure, if participants choose to withdraw, how will you deal with compensation?

There is no compensation involved for participants. If a participant asks to withdraw their data from the research project, all data correlated with that participant will be permanently deleted. This is stated in the Research Study included below, which will be signed by the investigators and participants prior to acquiring data from the participant.

CONSENT PROCESS:

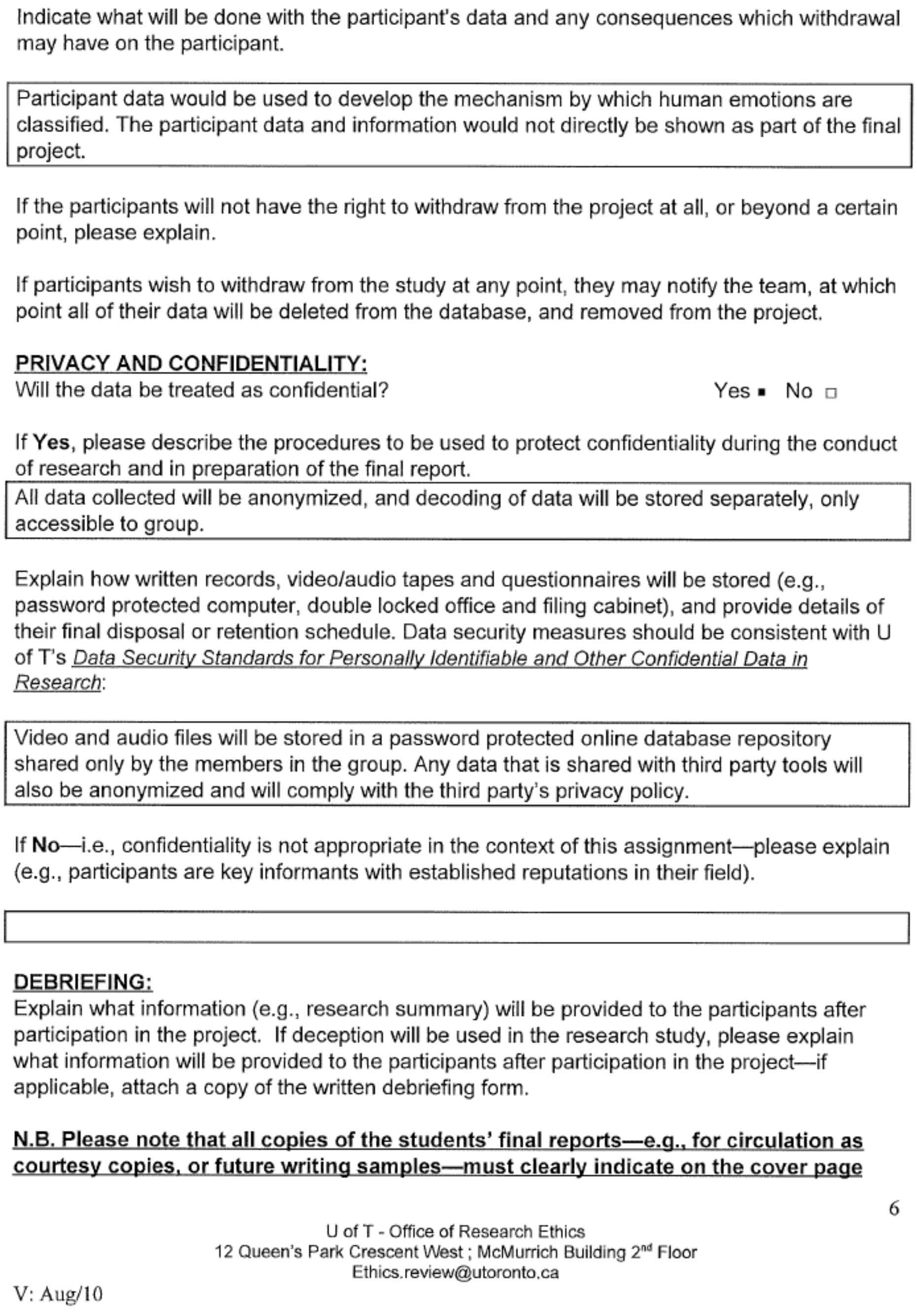
Describe the process that the student will use to obtain informed consent. Please note, it is the quality of the consent not the format that is important: if there will be no written consent form, please explain (e.g., if culturally inappropriate). If the research involves extraction or collection of personal information from a data subject, please describe how consent from the individuals or authorization from the custodian will be obtained. For information about the required elements in the information letter and consent form, please refer to: http://www.research.utoronto.ca/wp-content/uploads/2010/01/GUIDE-FOR-INFORMED-CONSENT-April-2010.pdf

N.B. Where applicable, please attach a copy of the Information Letter/Consent Form, the content of any telephone script, letters of administrative consent or authorization and/or any other material which will be used in the informed consent process.

If the participants are children, or are not competent to consent, describe the proposed alternate source of consent, including any permission/information letter to be provided to the person(s) providing the alternate consent as well as the assent process for participants.

All participants would be consenting legal adults.

Where applicable, please describe how the participants will be informed of their right to withdraw from the project. Outline the procedures which will be followed to allow them to exercise this right.



Indicate what will be done with the participant's data and any consequences which withdrawal may have on the participant.

Participant data would be used to develop the mechanism by which human emotions are classified. The participant data and information would not directly be shown as part of the final project.

If the participants will not have the right to withdraw from the project at all, or beyond a certain point, please explain.

If participants wish to withdraw from the study at any point, they may notify the team, at which point all of their data will be deleted from the database, and removed from the project.

PRIVACY AND CONFIDENTIALITY:

Will the data be treated as confidential? Yes ■ No □

If **Yes**, please describe the procedures to be used to protect confidentiality during the conduct of research and in preparation of the final report.

All data collected will be anonymized, and decoding of data will be stored separately, only accessible to group.

Explain how written records, video/audio tapes and questionnaires will be stored (e.g., password protected computer, double locked office and filing cabinet), and provide details of their final disposal or retention schedule. Data security measures should be consistent with U of T's *Data Security Standards for Personally Identifiable and Other Confidential Data in Research*:

Video and audio files will be stored in a password protected online database repository shared only by the members in the group. Any data that is shared with third party tools will also be anonymized and will comply with the third party's privacy policy.

If **No**—i.e., confidentiality is not appropriate in the context of this assignment—please explain (e.g., participants are key informants with established reputations in their field).

DEBRIEFING:

Explain what information (e.g., research summary) will be provided to the participants after participation in the project. If deception will be used in the research study, please explain what information will be provided to the participants after participation in the project—if applicable, attach a copy of the written debriefing form.

N.B. Please note that all copies of the students' final reports—e.g., for circulation as courtesy copies, or future writing samples—must clearly indicate on the cover page

the instructor, course number, and department or program at the University of Toronto that the report was prepared for.

Participants will be provided with a research summary detailing the purpose of the project, and how their data will be used to aid in the project development. See attached Research Summary. No deception will be used in the research study.

SIGNATURES:

As the **Principal Investigator** on this project, my signature testifies that I will ensure that all procedures performed under the project will be conducted in accordance with all relevant University, provincial and national policies and regulations that govern research involving human participants. Any deviation from the project as originally approved will be submitted to the Research Ethics Board for approval prior to its implementation.

Signature of Principal Investigator: Date: Nov. 17 2019

As the **Faculty Supervisor** on this project, my signature testifies that I have reviewed and approve the scholarly merit of the research project and this ethics protocol submission. I will provide the necessary supervision to the student researcher throughout the project, to ensure that all procedures performed under the research project will be conducted in accordance with University, provincial and national policies and regulations that govern research involving human subjects. This includes ensuring that the level of risk inherent to the project is managed by the level of research experience that the student has, combined with the extent of oversight that will be provided by the Faculty Supervisor and/or On-site Supervisor.

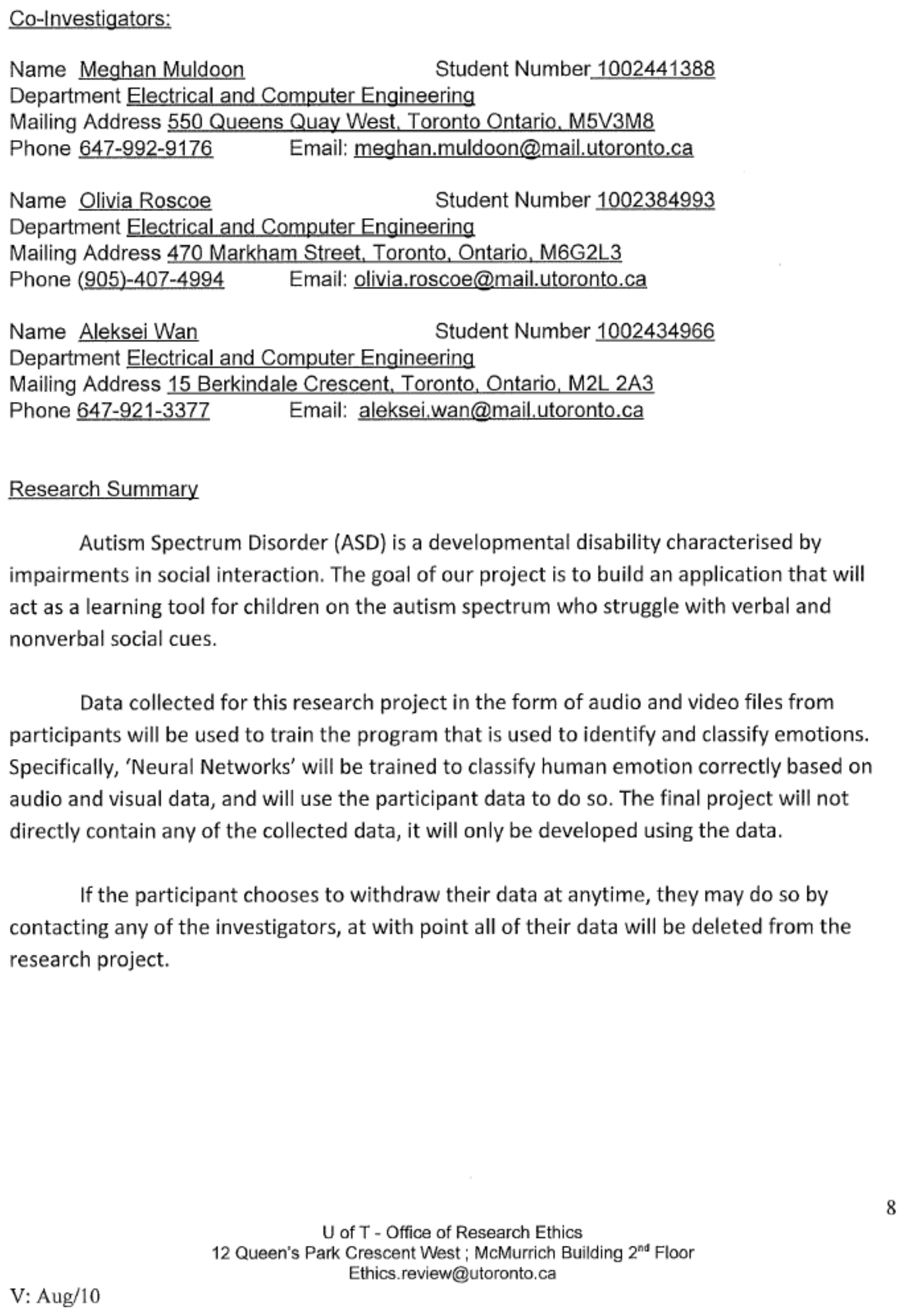
Signature of Faculty Supervisor: Date: Nov 22, 2019

As the **Undergraduate Coordinator**, my signature testifies that I am aware of the proposed activity, and understand that the level of risk inherent to the project should be managed by the level of research experience that the student has, combined with the extent of oversight that will be provided by the Faculty Supervisor and/or On-site Supervisor.

Signature of Undergraduate Coordinator: _ Date: Nov 27, 2019

As the **Departmental Chair/Dean**, my signature testifies that I am aware of the proposed activity, will allocate space and other resources required, and will provide administrative support to the research activity. My department, faculty or division will oversee the conduct of research involving human subjects to ensure compliance with University, provincial and national policies and regulations. My signature also reflects the willingness of the department, faculty or division to administer the research funds, if there are any, in accordance with University, regulatory agency and sponsor agency policies.

Signature of Departmental Chair/Dean: _ Date: NOV. 28, 2019



Co-Investigators:

Name Meghan Muldoon Student Number 1002441388
Department Electrical and Computer Engineering
Mailing Address 550 Queens Quay West, Toronto Ontario, M5V3M8
Phone 647-992-9176 Email: meghan.muldoon@mail.utoronto.ca

Name Olivia Roscoe Student Number 1002384993
Department Electrical and Computer Engineering
Mailing Address 470 Markham Street, Toronto, Ontario, M6G2L3
Phone (905)-407-4994 Email: olivia.roscoe@mail.utoronto.ca

Name Aleksei Wan Student Number 1002434966
Department Electrical and Computer Engineering
Mailing Address 15 Berkindale Crescent, Toronto, Ontario, M2L 2A3
Phone 647-921-3377 Email: aleksei.wan@mail.utoronto.ca

Research Summary

Autism Spectrum Disorder (ASD) is a developmental disability characterised by impairments in social interaction. The goal of our project is to build an application that will act as a learning tool for children on the autism spectrum who struggle with verbal and nonverbal social cues.

Data collected for this research project in the form of audio and video files from participants will be used to train the program that is used to identify and classify emotions. Specifically, 'Neural Networks' will be trained to classify human emotion correctly based on audio and visual data, and will use the participant data to do so. The final project will not directly contain any of the collected data, it will only be developed using the data.

If the participant chooses to withdraw their data at anytime, they may do so by contacting any of the investigators, at with point all of their data will be deleted from the research project.