Machine Learning Framework For Predicting Empathy Using Eye Tracking and Facial Expressions

MSc Research Project
Data Analytics

Samarth Krishna Dhawan
Student ID: 20180489

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahni
Supervisor: Dr. Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Samarth Krishna Dhawan
Student ID:	20180489
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Anu Sahni
Supervisor:	Dr. Paul Stynes
Submission Due Date:	15/08/2022
Project Title:	Machine Learning Framework For Predicting Empathy Using Eye Tracking and Facial Expressions
Word Count:	5709
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Samarth Krishna Dhawan
Date:	19th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Machine Learning Framework For Predicting Empathy Using Eye Tracking and Facial Expressions

Samarth Krishna Dhawan
20180489

Abstract

Empathy is one of the most important human emotion that facilitates a bond. This research provides a novel machine learning framework that combines eye tracking and facial expressions to predict empathy. The objective of this research is to help with recruitment of highly empathetic people in the medical and psychology domain for more empahtetic nurses and therapists. Features used were heatmaps from eye tracking and emotion detection from facial expressions were extracted which were given as inputs along age, sex, memory test score, blink percentage, blink mean, blink standard deviation, saccade percentage, saccade mean, saccade, average distance from left eye and average distance from right eye standard deviation to 3 machine learning models; Random Forest, Gradient Boosting and Logistic Regression. Logistic regression outperformed the other models with an F1-score of 0.74 while, Gradient boosting was the worst performing model with an F1-score of 0.4 and Random Forest had an F1-score of 0.5. YOLOv5 and Principal Component Analysis were also used to prepare the data and extract the right features for this model.

1 Introduction

In his study, Decety (2015) defines empathy as an intrinsic ability to notice and be sensitive to the emotional states of others paired with a willingness to care for their well-being. It is a crucial human feeling that facilitates and improves communication. Empathy can be divided into two broad categories, cognitive empathy and affective empathy. Cognitive empathy pertains to understanding the other person's state of mind and feelings, while affective empathy is more to do with the ability to respond with suitable emotion to others' emotional state. Affective and cognitive empathy are also distinct; someone who empathizes strongly emotionally is not necessarily good at understanding another's point of view. Both empathy quotients are commonly measured through various self-reporting questionnaires.

The aim of this research is to investigate to what extent can a combination of eye gaze patterns, facial expressions, memory test and demographic details can help in predicting empathy quotient in a person?

The eye gaze data was tracked using SMI equipment and facial expression data was tracked through a standard laptop webcam. YOLOv5 and Principal Component Analysis used as part of the machine learning framework to come up with the eventual empathy prediction.

There are particular professions that require recruitment of highly empathetic individuals. In the medical field, these professions include nurses and doctors. As it has been shown in previous research that if a patient feels comfortable and the doctor and nurse are able to communicate with them in an empathetic manner, the patient is more likely to recover quickly than when empathy is taken out of the equation Jongerius et al. (2021) & Aoki and Katayama (2021).

Additionally, empathy is absolutely imperative in psychology and psychotherapy. It enables in building a better counsellor-client relationship which aids effective treatment and interventions which results in better outcomes Wu et al. (2021). In a research Mawani and Nderu (2020) an online application was built which provided the least possible first-level counselling support with empathy at its core. This resulted in more such people actually going beyond the first-level of counselling to more face-to-face counselling sessions which helped them diagnosis and recovery from illness much quicker.

As discussed earlier in the section, both cognitive and affective empathy are usually measured through self reporting questionnaires. It has been previously found that self reporting questionnaires are not always the best way to measure this because not all people are fully aware of their cognitive empathy capabilities Dang et al. (2020). Efforts have been made to measure empathy using behavioural tasks, eye tracking etc. However, It is believed that facial expressions give even more insight into a persons empathy scoreLy and Weary (2021). Therefore, this research provides a novel machine learning architecture which analysis multiple cues and features like facial expressions, eye tracking, demographic information and memory test to predict empathy.

Because this is an interdisciplinary topic including both data analytics and psychology, a good grasp of psychology in terms of anticipating human emotions and, consequently, empathy is required. The literature review mentioned in the next part contributes to closing that gap.In addition, extensive study and literature analysis will be necessary to comprehend cutting-edge approaches and algorithms for ingesting data from eye tracking and facial expressions into the machine learning pipeline. The literature research undertaken in the following sections has offered a much better knowledge of emotions within the psychology area as well as several techniques to pre-process and forecast utilizing face and eye tracking information.

This paper discusses related work in section 2 to enhance the understanding of psychology and machine learning domains and to provide background on the latest research carried out in this particular area of empathy prediction through multiple cues. Section 3 presents the research methodology and framework used to carry out this research. Results are summarized and presented in section 4 while section 5 contains conclusion and scope for future work

2 Literature Review

Traditionally, most widely used way of measuring empathy has been through questionnaires using a likert scale. The following subsections will critically analyze the existing methods related to different parts of the project. The subsections include existing methods of empathy measurement, empathy using eye tracking, stimuli to invoke empathetic response, facial expressions and empathy and machine learning in empathy prediction.

2.1 Existing Traditional Methods of Empathy Measurement

Questionnaires have long been the go-to method of empathy measurement in the psychology research community. The questions are structured in a way where the participants are given certain situational questions and they are supposed to be answered on a likert scale. The likert scale can range from 3 point likert scale to a more elaborate 7 point scale. For example, the Toronto Empathy Questionnaire(TEQ) Spreng* et al. (2003) ranges on a 5 point scale (”Never”,”Rarely”,”Sometimes”,”often” and ”Always”)

One of the most popular and oldest questionnaire is the Interpersonal Reactivity Index Davis (1983) which was introduced in 1983 to measure empathy among other traits. It has 28 questions and have proven to work well for empathy measurement in nurses Aoki and Katayama (2021). It describes empathy as a multi faceted construct instead of a single point measurement. However, as this questionnaire is relatively old it was not suitable to be used directly as part of this research. Even though an update was made to IRI by polish academics Kaźmierczak and Karasiewicz (2021), it was still only tailored to find empathy among couples which doesn't align with the objective of this research it tries to predict empathy in individuals.

Another old yet popular questionnaire is the Impulsiveness-Venturesomness-Empathy (IVE-7) Eysenck and Eysenck (1978) which was designed in 1978. The questionnaire has also been used extensively, however, the purpose of the questionnaire was to find personality of a person as a whole rather than isolating the empathy quotient.

A common method observed in literature has been to modify or tweak existing traditional questionnaires to particular experiments to get the most accurate results. For example, a korean version of the popular Toronto Empathy Questionnaire(TEQ) Yeo and Kim (2021) was designed for medical students. It is imperative to match the questionnaire with the research objectives as shown in Stosic et al. (2022). The author argues that even though there are multiple questionnaire, the results from each of them are not correlated at all. This finding makes picking the right questionnaire for a particular research question absolutely imperative.

The research conducted by Olderbak et al. (2014) found that a person might not be equally empathetic towards all emotions. For example, a persons empathy quotient towards sadness might be higher than happiness depending on their state of mind. Hence the authors developed seperate questionnaires for different emotions. Based on finding this particular research will be using the sadness specific empathy questionnaire because the video stimuli being used is that of sadness.

2.2 Measuring Empathy Using Eye Tracking

Researchers over the years that have studied and correlated eye gaze behaviour with emotion have followed the same basic construct, Wherein a participant is shown some sort of video or pictorial stimuli while their eye gaze is tracked. Post which, a statistical study is conducted to find correlate eye gaze patterns with empathy scores. Eye tracking has gained traction over the last few years due to the availability of hardware and software that can produce eye tracking patterns along with other features link blinks, saccades, area of interest etc.Skaramagkas et al. (2021) Lim et al. (2020) Savin et al. (2022) Martinez-Marquez et al. (2021)

In research Cowan et al. (2014) the authors concluded that participants that had a high empathy score on self reported questionnaires tend to concentrate on eyes and face of the person in the video as against less empathetic participants who concentrated more on other regions of the video. This pattern was also noticed in an earlier research conducted by Decety Claus (2006). Hence, this gives us an indication that eye gaze patterns can be a helpful predictor of empathy.

A recent study Warnell et al. (2021) adding on to the work done by Cowan et al. (2014) furthered the evidence of highly empathetic people concentrating in the eye region of the narrators. However, the video stimuli used in their video was socially rich with multiple actors in a particular scene. Such a video stimuli has more probability of distracting the participants from the actual story and will provide a more scattered eye gaze patterns even for highly empathetic participants.

In another study Zaki et al. (2008) researchers tried the empathy accuracy task where in participants have to continuously rate the narrators perceived strength of emotion which is then compared against the narrators self reported strength of emotion.

2.3 Stimuli To Invoke Empathy

As discovered in the previous subsection, academics use a visual stimuli to generate an emotional response among participants. This subsection will explore various stimuli in greater detail. In Tarnowski Pawe (2020) and Cotler et al. (2020) researchers used video fragments from the famous movie forrest gump where the the actor is being bullied. However, the major drawback of such a stimuli is that the response of the participants who have already watched the video might not be authentic or as strong as it would have been on viewing the clip for the first time. Moreover, this also has the problem of a socially rich clip with multiple characters making hard to focus on one narrator.

Studies have also been conducted with images as stimuli in place of videos. For example in Ziaei et al. (2022) and Harrison et al. (2007) researchers used images to invoke a response from the participants. The images are black and white that contain both males and females photographed with various facial expressions like happy, disgust, sad etc. In other cases pictures of scenes or groups have also been used as stimuli where the image shows a violent protest or a group of people laughing together. These image based stimuli have proved to be effective for some researchers. However, these images are not engrossing enough, whereas video forces a person into a more wholesome experience with both visual and auditory senses getting involved.

Another popular response invoking task is the Emotional-Accuracy-Task(EAT) as discussed in the previous section. However, seeing the drawbacks of the EAT, this study Cowan et al. (2014) designed a modified version of the EAT where the narrators are professional actors that narrate hypothetical personal stories while looking directly at the camera. This creates a more engaging experience for the viewer as this gives an impression of the narrator directly communicating with the viewer.

2.4 Facial Expression and Empathy

Facial expressions have always been a window into the true emotions of humans. These expressions are very prominent and have been used to depict different emotions even in research. There are various computer vision algorithms that have been developed to help predict emotion through facial expression with great success. Taking this to the next step which will help predict empathy is called facial mimicry. Facial mimicry can defined as a persons' ability to feel others emotions and replicating them in their own expressions Fischer and Hess (2017). Although, researchers have argued that people might only mimic the emotion of others if they seem friendly or they share some common traits Hess and Blairy (2001). However, this might not be true because a study conducted by Sato and Yoshikawa (2007) that people reacted with visible facial movement to a video with a lag of just 900 milliseconds. This was done using facial action coding system which have facial actual units which tracks the muscles on the face to identify various facial expressions.

In yet another research, Ly and Weary (2021) the authors showed that even if participants were shown publicly sourced video clips of animals undergoing fairly regular procedures like castration, they showed empathy towards the animals pain through facial mimicry. The study found that people who self-reported being higher on the scale of unpleasantness also showed empathy towards animals through stronger facial mimicry and expressions. These results go on to show that facial expressions and facial mimicry can be a good feature to predict empathy quotient in a person.

The hypothesis of facial mimicry even holds true for adolescents according the research conducted by Van der Graaff et al. (2016). The researchers showed empathy invoking videos to adolescents and found that even adolescents that reported higher empathy had stronger facial expressions and mimicry.

In a more recent study, Drimalla et al. (2019) attempts to demonstrate a link between cognitive and emotional empathy and facial imitation were made. This study's findings also revealed that an individual's empathy score has a substantial relationship with facial expression mimicking. However, the findings revealed a greater relationship between affective empathy than cognitive empathy.

The research by Barbieri et al. (2019) also gives more support to the argument of facial expressions being a good predictor of empathy where they processed facial expressions in their raw state as well as using landmarks and discovered that they can be helpful in predicting empathy.

2.5 Machine Learning in predicting empathy

Machine learning has become a major part of every industry over the last decade because of cheaper computation cost, higher computation power and great volumes of data. It has made its way into all industries and helps run key process and is part of various pipelines across major corporations.

In a research done by Mathur et al. (2021), the authors attempt to predict participant empathy from a human robot interaction. When the participant enters the same room as the robot, the machine asks about the individual and then tells the participant three unique emotional stories. Following each narrative, the participant is expected to complete a questionnaire that assesses empathy and serves as the model's ground truth. While the stories are being told, an external camera records the participants' facial reactions. During the modeling step, video frames are recovered and numerous characteristics such as eye gaze, face landmarks, facial activity units, and so on are extracted, yielding 709 visual features per video. This data is then pre-processed before machine learning algorithms like adaboost and SVMs were applied to achieve an accuracy of over 70%

In another study, López-Gil et al. (2016) the researchers use EEG data in combination with eye gaze data to guess empathy in individuals. Images from a dataset depicting various emotions are displayed to participants. eye gaze plots from an eye tracking device, as well as electrical signal data from EEG, are pre-processed and inserted into a various algorithms, ranging from linear models like logistic regression to more complex models like the multi-layered perceptron. All algorithms gave different results on subset of participants. For example, MLPs outperformed all other models in one group while Random Forest outperformed all other models in another.

In an attempt to build an empathy machine Kummer et al. (2012) Researchers used an embedded camera to try to follow the user's facial expressions. Once the emotion has been identified, the machine plays some music that is appropriate for the feeling, much like background music in a movie. The claim is that when background music is present, two persons interacting will begin to experience emotional convergence, or the other person will always be aware of the user's emotions since the background music reflects the user's emotional state. The authors created and tracked 66 locations around a person's face to perform Facial Emotion Recognition (FER) using OpenCv.

In this study Barbieri et al. (2019) the researchers used a multimodal approach for empathy prediction. They processed the raw videos of facial expression separately while also processing the facial expression videos using facial landmarks. These were later combined in the modelling stage to come up with empathy predictions.

2.6 Summary Of Literature Review

As seen from the literature review, there has not been research conducted on combining different cues to predict empathy. There have been approaches including only eye tracking or only including facial expressions. This research bridges the gap and combines multiple human cues for prediction of empathy through machine learning

3 Research Methodology

The research methodology from data collection, data extraction, data processing, EDA, feature extraction and modelling is illustrated in Fig.1

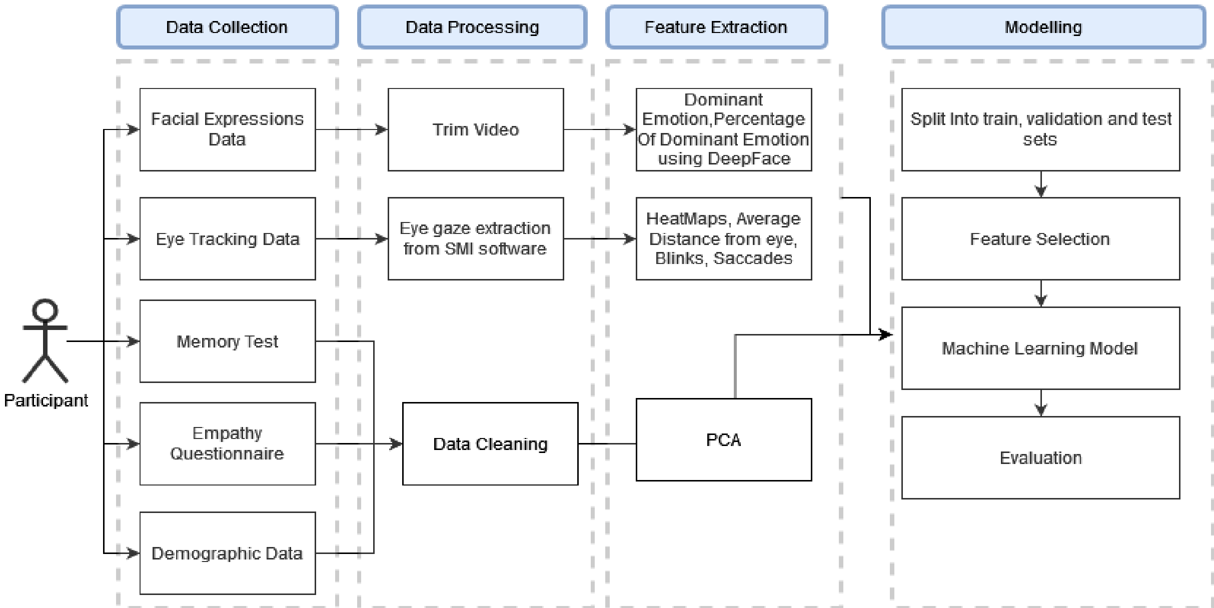


Figure 1: Research Methodology Overview

For data collection, 50 participants were invited and requested to watch a 13 minute video which features 6 different actors narrating personal hypothetical sad stories about their lives, while looking directly at the camera. These videos were sourced from Cowan et al. (2014) and Mackes et al. (2018) with prior approvals from the lead authors. Before starting the video, they were asked their levels of sadness on a scale of 1 to 10. While the participants were watching the video, there point of gaze was recorded using the eye tracking glasses by SMI and facial expressions were recorded through the webcam available on the same laptop. Once the participants were done watching the video, they were first asked the same question of level of sadness on a scale of 1 to 10 after watching the videos. This was done to get data on difference in self-reported sadness levels before and after the stimuli(Video). Post this, participants were requested to take a memory test based on the videos which consisted of 10 multiple choice questions about the stories in the stimuli. The last part of data collection was where participants filled out a sadness empathy questionnaire which had 10 questions on a 7-point likert scale measuring both cognitive and affective empathy("Disagree Strongly","Disagree Somewhat","Disagree Slightly","Neutral","Agree Slightly","Agree Somewhat","Agree Strongly").

In the data processing/data extraction phase, the eye gaze data was fed into a SMI's Bgaze software which produced the point of gaze. Along with the point of gaze it also produced some structured data in the form of blinks and saccades. While the facial expression videos were trimmed to match the start and end time of the experiment.

For feature extraction, to produce heatmaps from eye gaze data, two YOLOv5 models were built. The first model identified the eye gaze dot in the video while the other model identified where the laptop screen in the video was. This was done because the field of view of every person could be different and the laptop could be at a different angle and position for every participant which would make the scale of each of the heatmaps completely different. Once the models were trained, the coordinates of the bounding box for the eye gaze circle and laptop screen were combined to produce heatmaps for each of the participants. To calculate the average distance of the participants point of gaze from the narrators eyes, first the coordinates of the eyes of the narrator were extracted using the dlib library. Post this, using the bounding box prediction of eye gaze from the earlier YOLOv5 model, the average distance from both the left and right eyes were calculated using the simple euclidean distance measure.

Facial expressions data was fed into a pretrained DeepFace model, which produced the most dominant emotion in each frame of the video. Using this data, percentage of time the participant had sad as the most dominant emotion across the experiment was extracted. DeepFace is a deep learning model built by the research team at facebook and was trained on more than 4 million images to recognize faces and emotions with 97.53% accuracy.

Modelling this data involved first augmenting the heatmaps along with normalizing them. First various pre-trained CNNs among ResNet50, VGG16 were tried along with a custom CNN network as well. However, this failed to produce good results because of only 44 data points. In order to process heatmaps through classical machine learning pipelines, they were first flattened and PCA was applied on them to extract the features which explain 95% of the variance in the data which came to 34 features. Data was passed through various machine learning architectures which included all the variables namely Age, Sex, self reported sadness levels before and after experiment, difference between sadness levels, Memory test scores, Average distance from right eye, average distance from left eye, blink percentage, blink duration mean, blink duration standard deviation, saccade percentage, saccade duration mean, saccade duration standard deviation, percentage sad emotion.

(a) Laptop Screen Prediction Using YOLOv5 (b) Circle Prediction using YOLOv5

Figure 2: Example of Prediction Using YOLOv5

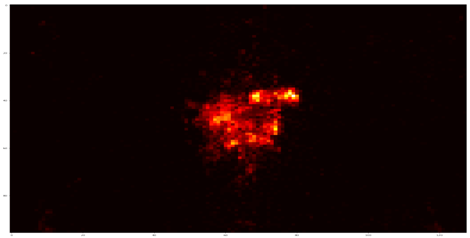


Figure 3: HeatMap Example

4 Design Specification

Before the experiment began, participants were asked to sign the consent form along with an excel sheet which asked them their Age, Sex and sadness levels on a scale of 1 to 10 before starting the experiment. This was done because it has been found in literature that demographic features are useful in empathy prediction. Sadness levels before were asked to ascertain the base sadness level of each participant which was later compared to any change in them after watching the stimuli video.

Post this, participants were made familiar with the eye tracking device and their point of gaze was calibrated in the SMI eye glasses. Data from eye tracking device is then processed through SMI Begaze software which helps in extract the video file with the point of gaze of the participant. Through the software, event statistics data which has information about saccades and blinks.

The facial expressions were recorded through the webcam embedded in laptop which was used for the experiment. This data was stored in cloud storage which was later downloaded to be processed in python. Once the facial expression data was downloaded, it was processed using DeepFace library which was developed by facebook and has an accuracy of 97.5% in recognizing emotions as it was trained on over 6 million images uploaded by facebook users.

The stimuli video was 13 minutes long which contains 6 actors narrating various personal sad stories while looking directly at the camera. For ease of participants to remember the stories, the narrators were given fictional names which were displayed for 10 seconds before the actor started narrating their stories.

After the video, participants were requested to fill out a memory questionnaire which consisted of the following 10 questions with 1 mark awarded for every right answer with no negative marking:

- Which relative of Bruce was suffering?
- What happened to Bruce's relative (disease)?
- What was stolen from Bruce?
- Why was Bruce sad after the item was stolen from him?
- To which position Selena was applying for?

- What was the actual reason that Selena didn't get the job?
- Which relative of Emma was suffering?
- What disease was Emma's relative suffering from?
- Which pet did Robert have?
- How does Robert describe his friend's sister?

This was followed by an empathy questionnaire which consisted of 10 questions and the options for each of the questions were "Disagree Strongly","Disagree Somewhat","Disagree Slightly","Neutral","Agree Slightly","Agree Somewhat","Agree Strongly". The marks for each of the options range from -3 to 3 with -3 indicating lack of empathy and 3 indicating the highest level of empathy while 0 was equivalent to a neutral emotion. The questions from the empathy questionnaire are listed below :

- I easily feel sad when the people around me feel sad.
- If a friend told me about an event in his/her life that made him/her feel sad, I will easily feel sad as well.
- I feel sad when I see that something is happening to a stranger that makes him/her feel sad.
- When I see that my friend is sad about something, I easily feel sad as well
- I am not easily infected by sadness of other people
- It is easy for me to understand why others become sad when something heartbreaking happens to them.
- It is difficult for me to understand what makes my friends sad
- I can easily think about events that will make my friends sad
- I have a hard time predicting what situations will make other persons sad.
- If someone tells me about an event that made him/her sad, I can easily understand why that event made him/her sad

After this, the participants were requested to update their self reported level of sadness between 1 and 10.

5 Implementation

The final implementation of the project is through a python based machine learning pipeline with some dependence on the SMI software and Questionnaire data as depicted in Fig.4

The implementation pipeline begun with pre-experiment questions of Age, Gender and sadness levels before the experiment These were recorded in a simple excel sheet. Post which the user watches the video. after watching the video, the user responded to the Memory questionnaire and Empathy questionnaire which were built using Microsoft forms. The answers of these are recorded in an excel sheet in the cloud and empathy quotient is calculated by taking the average of all answers as discussed in the previous section. This data was stored in a central database with a unique identifier for each row.

The eye tracking data was then fed into SMI Begaze software which produced the point of gaze. All the data including questionnaire, facial expressions and eye gaze data was processed through a python module. The questionnaire data was processed and new feature of the difference between sadness levels before and after watching the video was calculated. The Facial data was then processed using DeepFace library which was used to output the percentage of time the participant was sad during the video. The eye tracking data was processed using YOLOv5 weights to identify bounding boxes around the point of gaze and the laptop screen which was used to produce heatmaps. Confidence level for bounding box of the eye circle detection was set at 0.4 while that of laptop detection was set at 0.6. The mean average precision of both YOLOv5 models was over 0.995 and were trained for 300 epochs with a batch size of 8.

This data then traversed through the Pre-trained weights of a machine learning model which produced the final output in terms of a binary classification whether the person is empathetic or not.The entire pipeline is coded in python and the questionnaire data was stored on Microsoft Cloud.

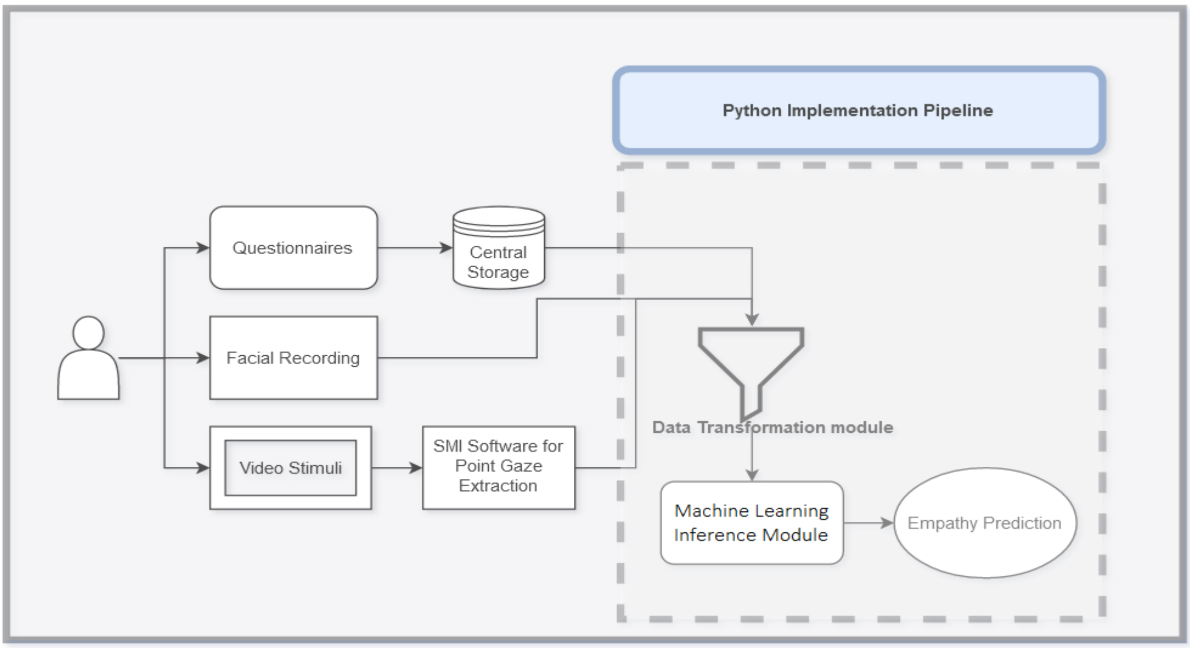
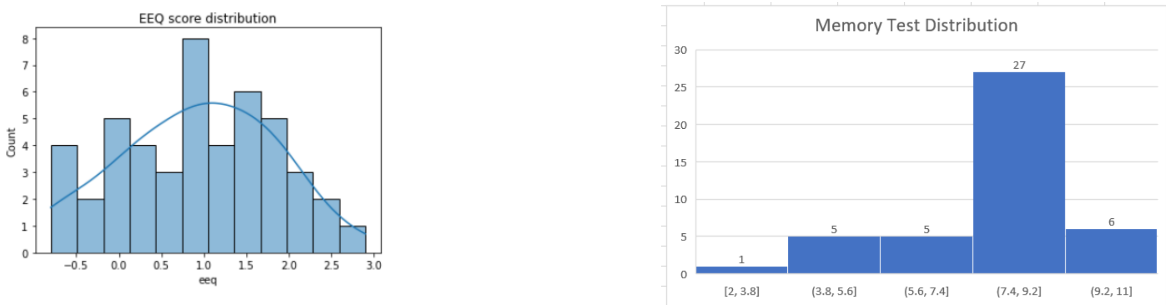


Figure 4: Design Implementation

6 Evaluation

The final data set consisted of 44 data points. Even though 50 experiments were conducted. 1 participant did not give consent to record their facial expressions. While the data for the other 5 could not be captured due to technical issues with the webcam. Hence,



(a) Empathy Scores

(b) Memory Scores Distribution

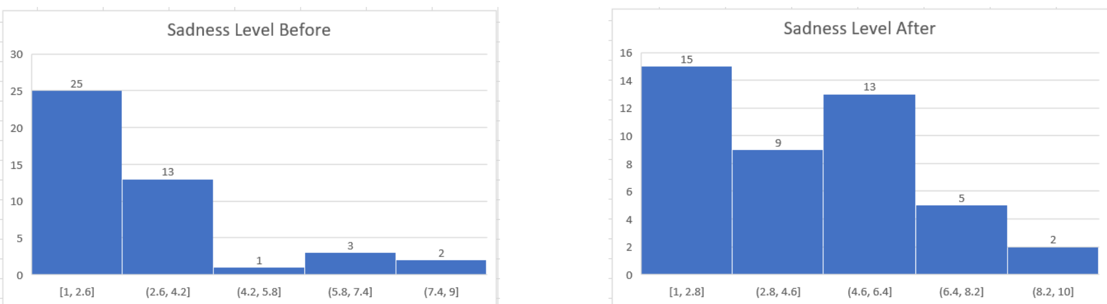
Figure 5: EDA of questionnaires

data of these 6 participants was completely removed before beginning the modelling exercise.

The data from empathy questionnaire was analyzed and it was found that the highest number of participants had an empathy score between 0.8 to 1.05 (8 participants) . Using this analysis the threshold for highly empathetic individuals was set at 1.25 which gave the data split of 17 highly empathetic people and 27 others.

Data from self reported sadness levels before and after the experiment revealed that the average sadness level increased by 62% from 2.5 to 4.06. Also, the correlation between EEQ and Difference of sadness levels after and before the experiment was 0.30. Participants in general scored highly in the memory questionnaire with the average score being 8 with only 11% of the participants scoring lower than 7 marks.

The performance metric that was optimized for in the pipeline was F1-Score as it takes both precision and recall into consideration and simple accuracy will not be a good metric choice for the model as the data is slightly imbalanced with an approximately 40-60 split. All models were built using a k-fold cross validation of 3 and Random Search was used to tune the hyper-parameters of each of the models.



(a) Sadness Scores Before Experiment

(b) Sadness Scores After Experiment

Figure 6: EDA of Sadness Levels

6.1 Experiment 1 - Logistic Regression with Only Facial Expressions

As part of this experiment, a simple logistic regression classifier was trained and the following hyperparameters were tuned using a random search

- C - regularization parameter that controls over-fitting. Tuned value of C - 0.0059
- Solver - this helps chose the algorithm that solves the underlying optimization problem. Tuned value of solver - saga
- Penalty - This helps decide which penalty to apply for regularization. Tuned value of penalty - l2

However, the model failed to perform well because the data has only one feature from facial expressions. The model reported an accuracy of 57% on the test set while the F1 score was 0, indicating that facial expression alone is not a useful feature in predicting empathy

6.2 Experiment 2 - Logistic Regression with only self reported features

In this experiment, a logistic regression was trained considering only the self reported features which are age, sex, sadness levels of the person before and after the experiment, difference between sadness levels and memory test scores. The model was tuned using random search with 3 cross validation sets.

- C - regularization parameter that controls over-fitting. Tuned value of C - 56.73
- Solver - this helps chose the algorithm that solves the underlying optimization problem. Tuned value of solver - sag
- Penalty - This helps decide which penalty to apply for regularization. Tuned value of penalty - l2

This model achieved an accuracy of 78.6% on the test set and an F1 score of 0.28

6.3 Experiment 3 - Logistic Regression with eye tracking features

for this experiment only eye tracking features were selected. These include, flattening the heatmap and doing a PCA to select 34 components that explain 95% of the variance. Other features include blink percentage, blink mean, blink standard deviation, saccade percentage, saccade mean, saccade standard deviation and average distance from left eye and average distance from right eye. The model was tune using random search for 3 cross validation sets

- C - regularization parameter that controls over-fitting. Tuned value of C - 2.28
- Solver - this helps chose the algorithm that solves the underlying optimization problem. Tuned value of solver - saga

- Penalty - This helps decide which penalty to apply for regularization. Tuned value of penalty - l2

The accuracy on the test set from this iteration was 67.5% while the F1 score was 0.66

6.4 Experiment 4 - Logistic Regression with all data

Since it was the objective of this research is to predict empathy using a combination of features hence, In this iteration, all feature sets were combined namely, facial expression features, eye tracking features and self reported features. Logistic regression using random search was tuned using 3 cross validation sets

- C - regularization parameter that controls over-fitting. Tuned value of C - 39.16
- Solver - this helps chose the algorithm that solves the underlying optimization problem. Tuned value of solver - liblinear
- Penalty - This helps decide which penalty to apply for regularization. Tuned value of penalty - l2

The model built on these features achieved the highest F1-score of 0.74 which emphasizes the importance of all features adding value in empathy prediction

6.5 Experiment 5 - Other Models on all data

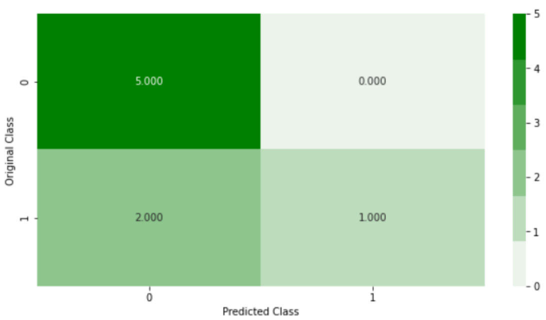
Random Forest and Gradient Boosting were the additional models built on combined data. Both models were tuned using random search on 3 cross validation sets. the hyper-paramters tune for Random forest were max depth, max features, min sample leaf, min sample split and n estimators. While the parameters tuned for gradient boosting were max depth, max features, min sample leaf, min sample split, n estimators and learning rate.

Random Forest had a precision and recall of 1 and 0.33 respectively while the F1-score was 0.5. Gradient boosting also had a recall of 0.33 while the precision stood at 0.5 giving it and F1-score of 0.4. Logistic regression was the best performing model of the three with recall and precision of 1 and 0.6 respectively giving it an F1-score of 0.74. The results can be seen from the confusion matrix in Fig.7

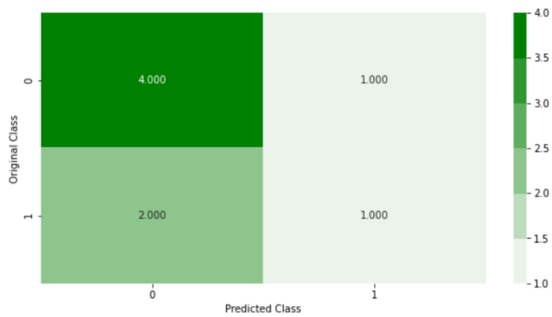
Artificial neural networks were also tried to generate predictions from images of heat maps, however, they failed to learn anything due to less data. The artificial neural network had 5 dense layers with relu activation and the last layer had softmax activation with binary cross entropy as the loss function and SGD as the optimizer.

6.6 Discussion

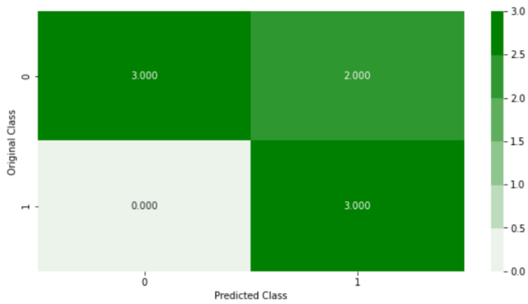
Even though individual features provided good accuracy; eye tracking features performed best with an F1-score of 0.66 while only facial expression feature did not add much value on its own. However, as hypothesised in the research earlier, combination of all features



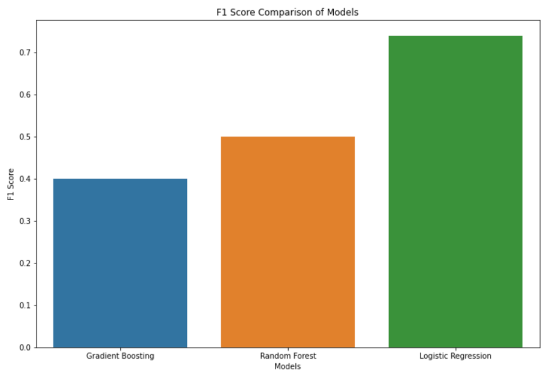
(a) Random Forest



(b) Gradient Boosting



(c) Logistic Regression



(d) F1 Score Comparison

Figure 7: Model Comparison With All Features

provided the best model with an F1-score of 0.75 and accuracy of 76% .

In the framework, facial expression features didn't work well on its own because of only one feature extracted from facial data. If more features were extracted, the experiments could have shown even better results. Also, memory test scores don't really help with prediction as almost everyone had a high score. However, it provides validation that participants were paying attention to the videos which gives validation that the experiments conducted are genuine and contain very little error.

PCA features extracted from heatmaps are some of the top features in variable importance graph shown in Fig.8 along with self reported sadness levels and average distance from eyes which shows that the hypothesis of correlation between empathy and eye tracking behaviour is true. Moreover, it can be seen from the same figures that even though emotion from facial expression is not the most important of features, it still adds value to the model.

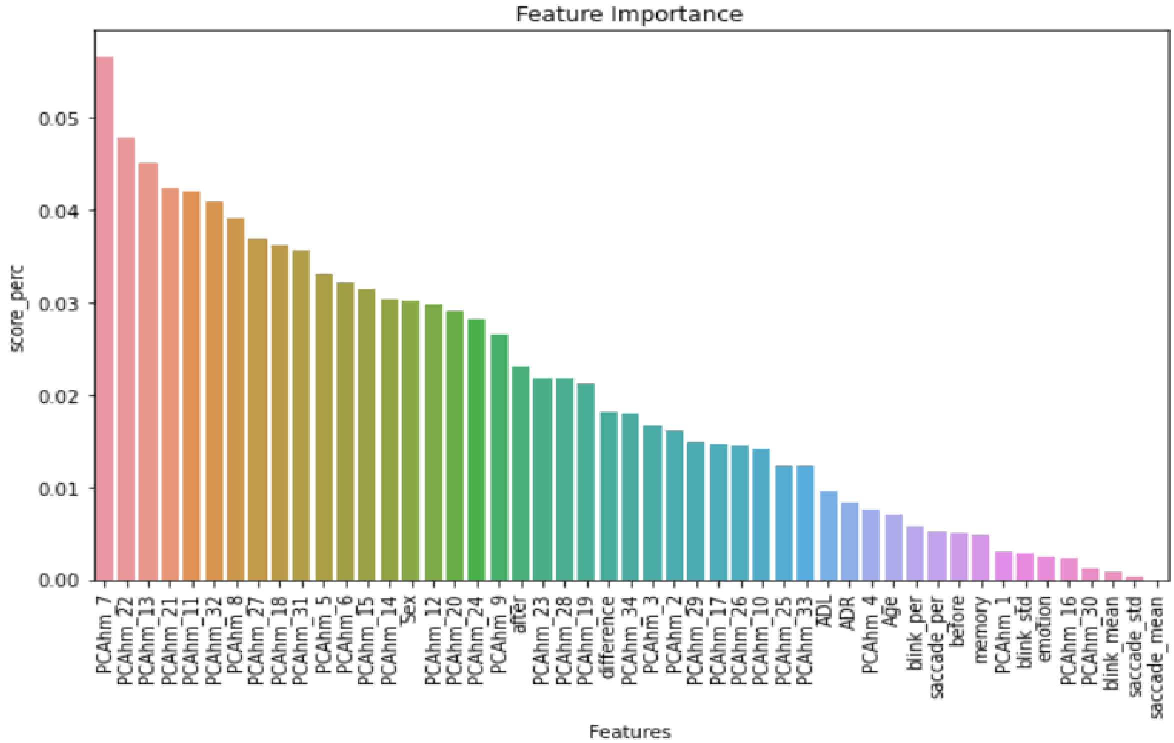


Figure 8: Feature Importance

Table 1: Comparison of Results

Feature Set	Model	F1 Score	Accuracy
Facial Expressions	Logistic Regression	0	57%
Self Reported Features	Logistic Regression	0.28	78.60%
Eye Tracking	Logistic Regression	0.66	67.50%
All Features	Logistic Regression	0.74	77%
All Features	Random Forest	0.5	75%
All Features	Gradient Boosting	0.4	63%

7 Conclusion and Future Work

The aim of this research was to provide a framework that combines multiple human cues; eye tracking and facial expressions that can help to predict empathy. It has shown that eye gaze patterns is a strong determinant of the empathy quotient in an individual. It also proves how emotion detection from facial expression adds value to the predictions. The results give support to the argument with the best model of logistic regression giving an F1-Score of 0.74.

As part of future work, this pipeline can be expanded by conducting the experiments on more participants with different visual stimuli. Also, more complex models especially to process heatmaps should be explored as this research has already shown the importance of eye gaze patterns in empathy prediction. CNNs, will be a good option to process heatmaps, which could not be successfully implement as part of this research due to lack of data. Also, other methods like landmark detection could be explored to increase the predictive accruacy of facial expression. If a more robust pipeline is built, it would drastically help in improving the predictive power of the models and thus help in recruiting of highly empathetic people in professions like nurses and therapists.

References

Aoki, Y. and Katayama, H. (2021). Development of the clinical interpersonal reactivity index to evaluate nurses' empathy, *Nursing & Health Sciences* **23**(4): 862–870.

Barbieri, F., Guizzo, E., Lucchesi, F., Maffei, G., del Prado Martín, F. M. and Weyde, T. (2019). Towards a multimodal time-based empathy prediction system, *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pp. 1–5.

Cotler, J. L., Villa, L., Burshteyn, D., Bult, Z., Grant, G., Tanski, M. and Parente, A. (2020). An interdisciplinary approach to detecting empathy through emotional analytics and eye tracking, *J. Comput. Sci. Coll.* **35**(8): 87–95.

Cowan, D. G., Vanman, E. J. and Nielsen, M. (2014). Motivated empathy: The mechanics of the empathic gaze, *Cognition and Emotion* **28**(8): 1522–1530. PMID: 24568562.

Dang, J., King, K. M. and Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated?, *Trends in cognitive sciences* **24**(4): 267–269.

Davis, M. H. (1983). Measuring individual differences in empathy: evidence for a multidimensional approach., *Journal of personality and social psychology* **44**(1): 113.

Decety Claus, J. L. (2006). Human empathy through the lens of social neuroscience, *TheScientificWorldJOURNAL* .

Decety, J. (2015). The neural pathways, development and functions of empathy, *Current Opinion in Behavioral Sciences* **3**: 1–6. Social behavior.

Drimalla, H., Landwehr, N., Hess, U. and Dziobek, I. (2019). From face to face: the contribution of facial mimicry to cognitive and emotional empathy, *Cognition and Emotion* **33**(8): 1672–1686. PMID: 30898024.

Eysenck, S. B. G. and Eysenck, H. J. (1978). Impulsiveness and venturesomeness: Their position in a dimensional system of personality description, *Psychological Reports* **43**(3_suppl): 1247–1255. PMID: 746091.

Fischer, A. and Hess, U. (2017). Mimicking emotions, *Current Opinion in Psychology* **17**: 151–155. Emotion.

Harrison, N. A., Wilson, C. E. and Critchley, H. D. (2007). Processing of observed pupil size modulates perception of sadness and predicts empathy., *Emotion* **7**(4): 724.

Hess, U. and Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy, *International Journal of Psychophysiology* **40**(2): 129–141.

Jongerius, C., Twisk, J. W., Romijn, J. A., Callemein, T., Goedemé, T., Smets, E. and Hillen, M. A. (2021). The influence of face gaze by physicians on patient trust: an observational study, *Journal of General Internal Medicine* pp. 1–7.

Kaźmierczak, M. and Karasiewicz, K. (2021). Dyadic empathy in polish samples: validation of the interpersonal reactivity index for couples, *Current Issues in Personality Psychology* **9**(4): 354–365.

Kummer, N., Kadish, D., Dulic, A. and Najjaran, H. (2012). The empathy machine, pp. 2265–2271.

Lim, J. Z., Mountstephens, J. and Teo, J. (2020). Emotion recognition using eye-tracking: Taxonomy, review and current challenges, *Sensors* **20**(8).

Ly, L. H. and Weary, D. M. (2021). Facial expression in humans as a measure of empathy towards farm animals in pain, *PLOS ONE* **16**(3): 1–16.

López-Gil, J.-M., Virgili-Gomá, J., Gil, R., Guilera, T., Batalla, I., Soler-González, J. and García, R. (2016). Method for improving eeg based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way, *Frontiers in Computational Neuroscience* **10**.

Mackes, N. K., Golm, D., O'Daly, O. G., Sarkar, S., Sonuga-Barke, E. J., Fairchild, G. and Mehta, M. A. (2018). Tracking emotions in the brain – revisiting the empathic accuracy task, *NeuroImage* **178**: 677–686.
URL: *https://www.sciencedirect.com/science/article/pii/S1053811918305068*

Martinez-Marquez, D., Pingali, S., Panuwatwanich, K., Stewart, R. A. and Mohamed, S. (2021). Application of eye tracking technology in aviation, maritime, and construction industries: a systematic review, *Sensors* **21**(13): 4289.

Mathur, L., Spitale, M., Xi, H., Li, J. and Matarić, M. J. (2021). Modeling user empathy elicited by a robot storyteller, pp. 1–8.

Mawani, A. and Nderu, L. (2020). Towards an online empathy assisted counselling web application, *EAI Endorsed Transactions on Context-aware Systems and Applications* **7**(22): 167792.

Olderbak, S., Sassenrath, C., Keller, J. and Wilhelm, O. (2014). An emotion-differentiated perspective on empathy with the emotion specific empathy questionnaire, *Frontiers in Psychology* **5**.

Sato, W. and Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions, *Cognition* **104**(1): 1–18.

Savin, G.-D., Fleșeriu, C. and Batrancea, L. (2022). Eye tracking and tourism research: A systematic literature review, *Journal of Vacation Marketing* **28**(3): 285–302.

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., Tripoliti, E. E., Marias, K., Fotiadis, D. I. and Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes, *IEEE Reviews in Biomedical Engineering* pp. 1–1.

Spreng*, R., McKinnon*, M., Mar, R. and Levine, B. (2003). The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures., *Journal of Personality Assessment* **91**(1): 62 – 71.

Stosic, M. D., Fultz, A. A., Brown, J. A. and Bernieri, F. J. (2022). What is your empathy scale not measuring? the convergent, discriminant, and predictive validity of five empathy scales, *The Journal of Social Psychology* **162**(1): 7–25.

Tarnowski Pawe, Kołodziej Marcin, M. A. R. R. J. (2020). Eye-tracking analysis for emotion recognition, *Computational Intelligence and Neuroscience* .

Van der Graaff, J., Meeus, W., de Wied, M., van Boxtel, A., van Lier, P. A. C., Koot, H. M. and Branje, S. (2016). Motor, affective and cognitive empathy in adolescence: Interrelations between facial electromyography and self-reported trait and state measures., *Cognition Emotion* **30**(4): 745 – 761.

Warnell, K. R., De La Cerda, C. and Frost, A. (2021). Disentangling relations between attention to the eyes and empathy., *Emotion* .

Wu, Z., Helaoui, R., Reforgiato Recupero, D. and Riboni, D. (2021). Towards low-resource real-time assessment of empathy in counselling, *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Association for Computational Linguistics, Online, pp. 204–216.
URL: *https://aclanthology.org/2021.clpsych-1.22*

Yeo, S. and Kim, K.-J. (2021). A validation study of the korean version of the toronto empathy questionnaire for the measurement of medical students' empathy, *BMC Medical Education* **21**(1): 1–8.

Zaki, J., Bolger, N. and Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy, *Psychological Science* **19**(4): 399–404. PMID: 18399894.

Ziaei, M., Oestreich, L., Persson, J., Reutens, D. C. and Ebner, N. C. (2022). Neural correlates of affective empathy in aging: A multimodal imaging and multivariate approach, *Aging, Neuropsychology, and Cognition* **29**(3): 577–598. PMID: 35156904.