National College of Ireland

Machine Learning Framework for Prediction of Empathy using Eye-tracking and Speech Analysis

MSc Research Project
Data Analytics

Rahul Badarinath
Student ID: 20247702

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:	Rahul Badarinath
Student ID:	20247702
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Anu Sahni
Submission Due Date:	15/08/2022
Project Title:	Machine Learning Framework for Prediction of Empathy using Eye-tracking and Speech Analysis
Word Count:	4540
Page Count:	21

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:	
Date:	15th August 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 Prediction of Empathy using Eye-tracking and Speech Analysis

Rahul Badarinath
20247702

Abstract

Empathy is often described as innate ability of an individual to perceive and sensitize to the emotional feelings of another individual and present a motivation to showcase care & affection to them. It is often believed, that empathy is the most important emotion towards building a better and more sustainable society. The objectives of this research are to help develop and improvise the current recruitment methodologies in the medical and psychological domains such as Nurses and Therapists, who require to be empathetic. The traditional approach for predicting empathy is through questionnaires. This research proposes a novel approach for predicting empathy using Machine Learning, Eye-tracking, and Speech Analysis by creating a robust framework. The framework comprises of numerous features such as Heatmaps, which were generated through Eye-tracking, most frequent emotion using Speech signals, Demographic details of a participant such as gender, age, memory test score, levels of sadness before and after watching the videos. Furthermore, metrics such as blink percentage, blink mean, blink standard deviation, saccade percentage, saccade mean, saccade, average distance from both eyes, standard deviation were also derived. A combination of these metrics were fed as inputs to three Machine Learning models; Gradient Boosting, Random Forest, and Logistic Regression.The YOLOv5 model and Principal Component Analysis were used to extract multiple features and in the preparation of data respectively. Logistic regression based model outperformed the rest with an F1 score of 0.86, Gradient Boosting a close second with an F1 score of 0.57, and Random Forest with an F1 score of 0.4..

1 Introduction

The authors of Decety (2015) describe Empathy as an innate ability of an individual to perceive and sensitize to the emotional feelings of another individual and present a motivation to showcase care & affection to them. This ability is regarded as the most important human emotion by various researchers in the psychological community, over which studies are being conducted from decades. Empathy is broadly classified into two categories, namely Cognitive and Affective Empathy. Cognitive Empathy emphasizes on the ability of an individual to transfer emotions on to oneself and imbibe the emotion as if they were their own. However, Affective Empathy is the ability of an individual to feel and understand the emotions relayed by another individual while their emotional state remains the same. The traditional approach to derive the Cognitive and Affective empathy of people is done through self reporting Questionnaires.

The following sections describes the merits and de-merits of Empathy alongside the research objectives and motivation.

The primary objective of this research is to effectively analyze to what extent the combination of eye-gaze patterns, speech signals, demographics of participants, and a memory-test can contribute towards predicting the empathy of a person?

The data for eye-gaze patterns was procured using SensoMotoronic Instruments (SMI) apparatus. Speech signals were recorded using a standard phone-based voice recorder. Principal Component Analysis and YOLOv5 were deployed to procure further metrics that further contributed in the design of the machine learning framework and the eventual empathy prediction.

This research bases emphasis on the importance of empathy in the fields of Healthcare, specifically with doctors and nurses. The authors of Jongerius et al. (2021) & Aoki and Katayama (2021) support this hypothesis by stating the needs for empathetic medical professionals. The patients are known to respond well to treatment when the doctors and nurses associated with the patient showcase empathy. This profession places a lot of emphasis on nurses to be empathetic, more so, as taking care of the patient post-op. Hence, the recruitment for such positions requires highly empathetic individuals.

The research also places emphasis on the need for empathy in the fields pertaining to Psychology and its therapeutic professions. The professionals with whom an individual shares their concerns requires the assessment of doctors who are not only psychologically trained, but those who have a high Affective Empathy quotient. The authors of Wu et al. (2021) concur with this analogy. The research conducted by the authors of Mawani and Nderu (2020) showed the improvement in the number of people who wished to attend counselling sessions. The experiment was performed by creating an online-based application where least amount of support was provided while maintaining empathy as the core-requirement at the first level to patients. Due to the emphasis based on empathy, the individuals felt encouraged to further attend counselling session, which helped in the diagnosis and prognosis, which eventually helped the individuals recover.

The research emphasizes on further understanding the need for cognitive and affective empathy. As discussed earlier, the scores for these were procured with the help of self-reported questionnaires. The authors of Dang et al. (2020) state the problem with this approach by stating the possibility of individuals not being capable of their cognitive empathetic capabilities. Hence, questionnaires may not be the best measure to depict empathy. The last decade has seen a rise in recognizing emotions and further empathy levels using facial expressions, and speech signals. The speech emotion recognition space has received a lot of traction. The authors of Casas et al. (2021) conducted an experiment where the patients in a psychology session were recorded and the speech signals were analyzed later to observe the difference in the auditory properties of an individual narrating a sad story vs the doctor who was required to maintain a neutral emotion throughout.

This research essentially fuses the world of Psychology and Data Science. This places emphasis on ensuring that the knowledge and domain understanding thus acquired is high to ensure effective and best results. The Literature Section below provides an in-depth understanding of concepts like Empathy and Speech, algorithms and techniques that can be deployed in the design of the machine learning framework, and the various methods through which the data can be ingested into the pipeline. This critical analysis of previous literature provides an understanding towards reaching the desired objectives outlined earlier.

The research is structured into multiple sections that helps in achieving the research

objectives. Section 2 provides critically analyzed literature providing an in-depth understanding of the current research, algorithms, and methodologies deployed in the Psychology and Data Science domains. Section 3 outlines the research methodology and the framework leveraged to conduct the research. Section 4 provides insights into the design and specifications leveraged to create the machine learning framework. Section 5 summarizes the results procured. Section 6 comprises of the conclusions and scope for further improvements.

2 Literature Review

As discussed in the previous section, the most widely used tool for measuring empathy is through questionnaires using a point-based likert scale. The following sub-sections provides literature on the existing methodologies in relation this study. These include traditional methodologies for measurement of empathy. measuring empathy using eye-tracking, measurement of empathy using speech analysis, speech emotion recognition using machine learning, visual stimuli for empathy prediction, summary of the literature review.

2.1 Traditional Methodologies for Measurement of Empathy

Empathy Questionnaires have been predominant in the psychology community where self-reporting empathy through questionnaires are highly encouraged, invoking the spirit of conducting research in this space as well. The questions portrayed in these questionnaires are often than most designed to be situational, where the solutions for these ranges typically from a 3-point to a 7-point likert scale. One such example is the Toronto Empathy Questionnaire (TEQ), designed by the authors of Spreng* et al. (2003) which consists of sixteen questions, with a 5-point likert scale, *"Never", "Rarely", "Sometimes", "Often", "Always"*.

The oldest questionnaire designed to measure empathy was the Interpersonal Reactivity Index (IRI), designed by the authors of Davis (1983). Introduced in 1983, this questionnaire had 28 questions, specifically designed to measure empathy, essentially performed well on nurses in the healthcare domain. Aoki and Katayama (2021). The authors described Empathy as a multi-faceted approach based measurement, and not a single-point measure, meaning the trait could not have been measure with just one parameter. However, this being introduced in 1983, meant we couldn't use it. The authors of Kaźmierczak and Karasiewicz (2021) created a modified study to use the questionnaire which was relatively recent. However, the study was only created for couples, which did not align with the research objectives.

Similarly, another questionnaire termed as "Impulsiveness-Venturesomness-Empathy" (IVE) designed by the authors of Eysenck and Eysenck (1978) was specifically designed to determine a person's personality traits, where empathy played a role, but wasn't the main objective. Despite the popularity of the use of this questionnaire in the psychology domain, the use of the same was rejected.

A commonly observed tactic or technique is the modification of existing questionnaires by tweaking some of the questions or the ranges of the likert scale or by simply changing the measure to procure accurate and precise outputs. One such study conducted by the authors of Yeo and Kim (2021) created a modified TEQ questionnaire primarily for students in the medical profession. The authors of Stosic et al. (2022) state the emphasis

on ensuring the questionnaire selected aligns with the research objectives. They postulate that results from different questionnaires are not inter-related and hence, mean different things in different contexts. This made choosing the right questionnaire that much more difficult and important for this research.

The authors of Olderbak et al. (2014) stated through their research that the empathy levels for a person would vary depending on the emotion. To elaborate, the authors say that a person's level of empathy towards happiness (wherein the person relates to happiness and wishes to share the emotion with another person) could be higher or lower than the same person's empathy for sadness as an emotion. Hence, emotion plays a major role in determining the person's empathy. The authors thus created a questionnaire which was emotion-specific. The questions are based on 6 emotions, Happy, Sadness, Anger, Disgust, Excited, and Surprised. Hence, the questions pertinent to Sadness will be used to conduct this research, since the visual stimuli for this research is pertinent to Sadness.

2.2 Measuring Empathy Using Eye Tracking

The fundamental construct that researchers adhere to finding or measuring the empathy of individuals using eye-tracking has remained the same. The participant is asked to view a series of visual stimuli, and the corresponding gaze pattern for the duration of the videos is procured. The stimuli could be video-based or image-based depending on the research Skaramagkas et al. (2021). On completion, the researchers analyze the data based on statistical studies, and map the gaze patters with the stimuli to obtain an emotion Lim et al. (2020). This methodology has recently come to light, and has since gained significant traction in the research communities. This traction is due to the advancements in technology that has provided capabilities to track the movement of eyes using various hardware and software, and provide other supportive metrics such as area of interests, saccades, pupil dilation, blinks, etc. Savin et al. (2022) Martinez-Marquez et al. (2021)

A research conducted by authors of Cowan et al. (2014) came to a significant conclusion that participants with high empathy through self-reported questionnaires would maintain their focus and eye-contact with the actor's eye's or face on screen. However, the less-empathetic people were presumed to concentrate but less on the face and more on other parts of the videos. This statement was also supported by the authors of Decety Claus (2006) and as they noticed a similar trend through their research. This indicates that using the position of the eyes through the point of gaze could be vital in determining the empathy of an individual.

Similarly, the authors of Warnell et al. (2021) also concluded with the same results as mentioned above. However their visual stimuli included multiple actors narrating a group conversation. In such cases, the possibilities of distraction is higher than that of concentration. Hence, the takeaway from this research is the fact that they are able to support the hypothesis outlines by Cowan et al. (2014) regarding the focus of the participant on the eyes of the actors on screen.

Another research conducted by the authors of Zaki et al. (2008) outlines a methodology where participants would be asked to rate their emotions continuously during the visual stimuli on screen with the mouse and a emotion scale provided at the bottom of the screen. On completion, the average emotion displayed by the participant throughout the process is used to determine the participant's level of empathy. However, this would hinder the objectives of this research as well as expecting a participant to continuously

scroll and rate themselves while watching the video could prove counter-productive for the research objectives.

In summary, the study conducted by Cowan et al. (2014) aligns the most with the research objectives and thus will be used to obtain the results.

2.3 Measurement of Empathy using Speech Analysis

Empathy is an emotional trait, where one human is capable to understand and share the feelings of another human. The measurement of empathy is proven to be beneficial across diverse fields, from healthcare to automotive, to schools and colleges.Nimmagadda et al. (2022).The research objectives are to design an alternative approach to detecting empathy in the recruitment of health-care individuals. The authors of Ooi et al. (2014) justify the same by showing the importance of speech in healthcare sector. The nurses are not just required to be empathetic by nature, but should also portray the same through the way they speak. This is essential because a patient can get easily offended if responses to their questions or fears is not met with an empathetic response. To ensure this, doctors are also requested to be empathetic while speaking to the patient or the family. This experiment done by the authors of Morais et al. (2022) discuss the application of detection of emotions in the healthcare sectors in detail. Speech signals can transmit meaningful data from humans rapidly. These signals are essentially used to detect the emotional state of the human beings. Speech is essentially comprised of seven prosodic features, "Accent", "Stress", "Rhythm", "Tone", "Pitch", and "Intonation" Ooi et al. (2014). Each of these features plays a vital role in determining the emotion behind a human.

The concept of understanding empathy is not just limited to humans. The expansion of the ability to adhere to the psychology behind emotions is underway for robots, and computers, who play a major role in the worlds of Human-Robot Interaction (HRI), and Human-Computer Interaction (HCI). The role of speech in this regard has been explored for decades now. However, a single feature cannot be used to detect an emotion due to an overlapping factor of the values of these features for one or more emotions.

2.4 Speech Emotion Recognition using Machine Learning

This section reviews literature on the various algorithms, classifiers, and methodologies adopted in detecting emotion through speech. This critical assessment, helped pave the way in applying the best approach for this research.

The most recent research done on SER by Singh (2022) suggests a novel methodology to predict emotions with the help of speech analysis by using Bi-directional long short-term memory cell (LSTM) architecture. It this research, the process that has been undertaken is, where the speech signal is obtained, and transfer-learning methodology is applied to obtain the Convolutional Neural Networks (CNN) features from a Resnet model, and these these features are trained on a bi-directional LSTM model and the confidence intervals or probabilities score is received as a result. These results looked promising, and the training accuracy was 82.03% , and the testing accuracy was almost 81% on the RAVDESS dataset.

Another research done by Nimmagadda et al. (2022) outlines the uses of humanoids in the health-care domain. A robot was deployed to a health-care facility to help mentally-ill patients. Mathur et al. (2021) They believed that the best way to teach a robot human emotions, is to combine facial expression data and voice samples of the patients at the

facility. Hence, the patients were constantly monitored , and then, 3 models based on Support Vector Machines (SVM), Multiple Regression, and Artificial Neural Networks (ANN) were deployed, and ANN, with the highest accuracy was used to build the final model, as mentioned in Přibil and Přibilová (2012), and Kammoun and Ellouze (2006).

A reseach conducted by Khalil et al. (2019) discussed the deployment of hidden and Gaussian Markov Models (HMM, and GMM), and other neural network architectures such as Recurrent Neural Networks (RNN), CNN, DCNN, etc are discussed in detail, for SER. The conclusion is that, while Deep learning is a very well-established way to work with labelled or unlabelled data, depending on the size of data, the model may out-perform or under-perform. Hence, care needs to be taken to ensure, while training the model, the features, number of parameters, and the speech data provided, does not hinder with the paradigm of analysis Schoneveld et al. (2021).

Finally, the approach provided by Chen et al. (2018) (Model 1) seems to be the best approach to detect emotions using speech. While the approach outlined in Singh (2022) (Model 2) is far more recent an novel, it observed the the research done by the authors of Chen et al. (2018) provided a better accuracy and more efficient model for deriving the emotions based on speech. The models in comparison are both based on Convolution Neural Networks, and are trained on the same RAVDESS dataset. However, the distinction is that Model 1 makes use of *Attention Networks*, while Model 2 uses *LSTM architecture* on top of the CNN deployed. The performance of Model 1 with an overall accuracy of approximately 88% was observed to be greater than that of Model 2 with an overall accuracy of 81%.

2.5 Visual Stimuli for Empathy Prediction

The use of visual stimuli is preferred by professors in academia as discussed in previous sections. The various stimuli preferred is reviewed in detail.

The authors of Tarnowski Pawe (2020) and Cotler et al. (2020) use of video frames from the movie *Forrest Gump* for predicting empathy. However, a major flaw in this approach is endangering or biasing the results. This is because the participants could've watched the videos prior to the experiment, as it is a famous movie. This taints the dataset because an authentic response is required from the students. The participant watching it again would not succumb to the actual emotional response intended to acquire. Moreover, the video of someone getting bullied, which was used contains the use of multiple characters in one scene, which would make the participant unable to focus.

Processing of videos for finalizing results is complex and time-consuming. To counter these problems, the authors of Ziaei et al. (2022) and Harrison et al. (2007) propose the use of imagery stimuli to depict empathy. These images used are pertinent to both Male and Female, which are Black and White in nature. These images are supposed to depict emotions such as Happiness, Sadness, Anger, Disgust, Excited, etc. Some of the reviewed papers also show the use of pictures involving a group of actors which is defining a group emotion such as a protest initiated fight or violence, or a laughter group of old men and women. However, these images cannot depict the empathy levels to the extent that this research aims to. Hence, the use of imagery based stimuli is rejected.

The Emotional-Accuracy Task (EAT) as discussed in the previous section is another popular methodology for invoking an empathetic response. There were a few drawbacks with the use of the EAT based approach, and hence, a modified study based on EAT was conducted by Cowan et al. (2014). In this study, professional actors are tasked to

enact sad-stories while directly looking at the camera where the stories are hypothetical or scripted. The fact that actors are narrating a story on camera whilst looking at it, creates an environment where the participant feels connected and engaged, thereby indulging and providing a much more substantial and accurate response.

2.6 Summary Of Literature Review

The critical literature reviewed proves the lack of research that has been conducted with a combination of eye-tracking, and speech towards depicting empathy. However, there exists research in predicting empathy using Eye-tracking and Speech individually. Hence, this research aims at combining these cues that would essentially assist in measuring the empathy levels of participants.

3 Research Methodology

The section comprises of the research methodology which includes the data gathering, data extraction and pre-processing, Exploratory Data Analysis, Feature extraction methods, and Modelling is illustrated in Fig.1

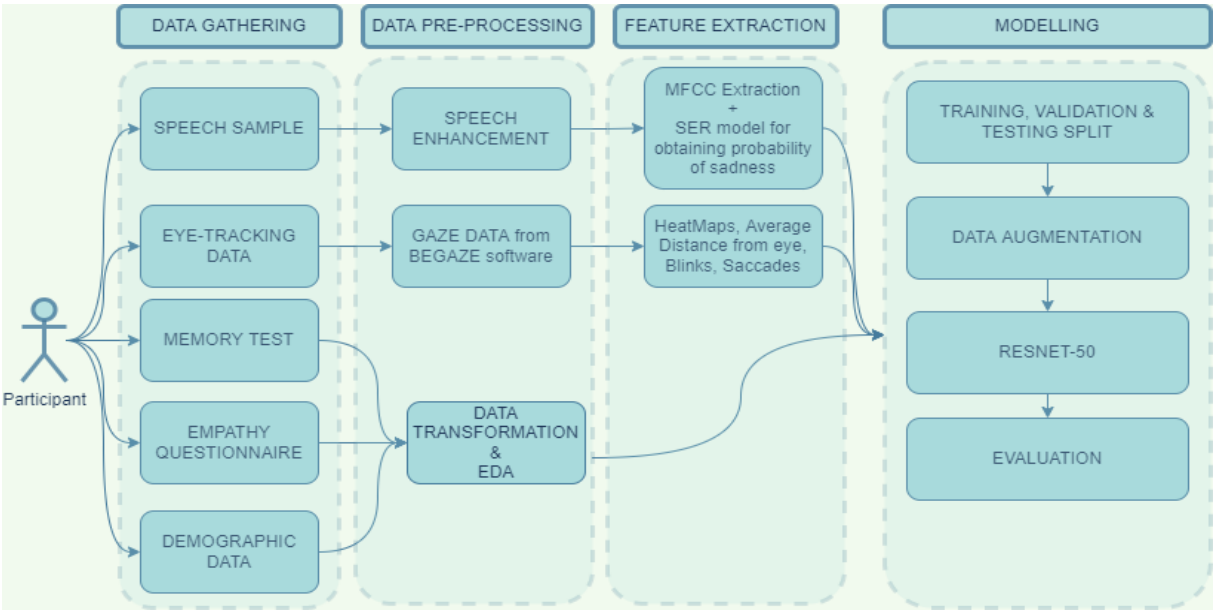


Figure 1: Research Methodology Work-Flow

The first phase is the data-gathering phase where data was collected from 50 participants, who were asked to view a video-based stimuli, 13 minutes in duration which consisted of 6 different actors expressing and narrating sad scripted stories from their lives. For the duration of the video, the actors made direct contact with the camera while they narrated the stories. The videos were procured as a secondary data source from Cowan et al. (2014) by performing the necessary data and ethics-related procedures. The participants' levels of sadness prior and post the videos were noted ranging between 1 and 10. The eye-gaze of these participants were recorded whilst watching the video using eye-tracking glasses designed by SensoMotornic Instruments (SMI). The speech data was recorded by the participant, on completion of the video where each participant was asked

to answer any one of the two questions, could they narrate a sad story from their life, or relate to one of the videos shown and why?

Finally, the participants were asked to take a memory test which consisted of 10 questions based on the videos that they watched to understand their level of focus and attention during the process, and finally to fill out a questionnaire based on empathy for sadness as an emotion, which comprised of 10 questions based on Cognitive & Affective Empathy levels, each with a 7 likert scale-ranging from *"Disagree Strongly","Disagree Somewhat","Disagree Slightly","Neutral","Agree Slightly","Agree Somewhat","Agree Strongly"*.

The next phase comprises of the data pre-processing carried out in this research. For the eye-gaze videos, these were fed into the SMI's Begaze Software through which the gaze corresponding to the participants were procured. Alongside, important metrics in respect to Visual Stimuli, Blinks, Saccades, AOI, etc were also procured in the form of a text document, that was then extracted into a excel sheet for processing. The speech samples that were recorded were fed into the Praat to clean, and re-sample at 16000 kHz, which was then passed into a Speech Enhancement model, where any background gaussian noise corresponding to the sample was removed using de-noising algorithms, and the cleaned speech signal procured was brought back into Praat for the next step.

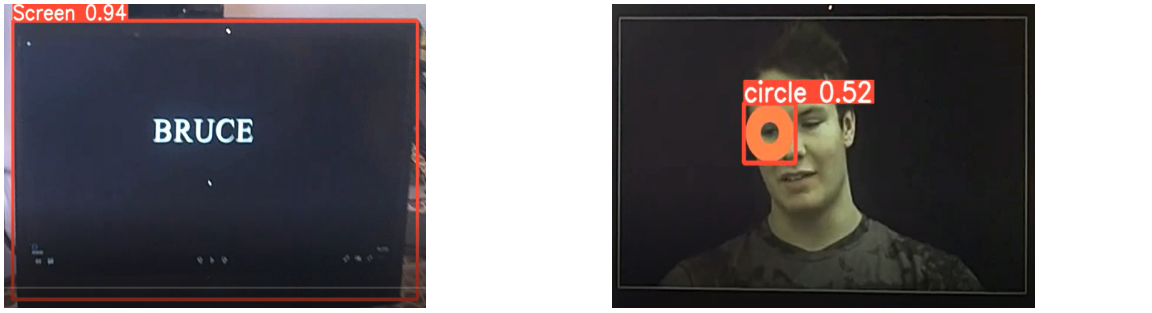
Further, the eye-tracking feature information pertinent to the research was divided into two parts. First part, eye-tracking feature extraction was done by creating heatmaps for each of the participant's video. For the same, two YOLOv5 models were designed. Of the two, the first was intended to to identify the eye-gaze "dot" in the video frame-by-frame as shown in Figure 2b and the second model was designed to identify the co-ordinates of the laptop screen where the visual stimuli was being displayed, as shown in Figure 2a. The first model provided clarity on where the participant was viewing within the screen at every ms of the video for the duration, while the second model provided specific and precise co-ordinate information on where the laptop screen was situated at every ms of the video, as the participant tends to move during the video, the position of the screen also changes in terms of angle of sight, and field of view. On completion of model building & training, the bounding boxes procured from each of these models were taken to produce the heatmap for each participant as shown in Figure 3. Furthermore, the co-ordinates from the bounding boxes of YOLOv5 model 1 was used to calculate the average distance between both eyes using *Euclidean distance* formula.

The speech samples procured from the previous step were the cleaned samples with minimal or no background noise. These samples were then fed into Praat, from where the Mel-Frequency Cepstral Coefficients (MFCC) features were extracted. These MFCC features attribute to 39 features for a speech sample, of which 12 parameters are related to the amplitude of frequencies, thereby producing enough frequency channels to analyze the audio sample. The rest of the features correspond to energy, pitch, logarithmic amplitude of the speech signal, and so on, which are primitive and important for this research. These features were extracted with a window length of 25 ms and a Hanning window of length 10ms. This data was then further simplified using logarithmic and average based measures to procure a single numerical value across the sample for analysis. The speech-sample was also fed into a pre-trained Speech Emotion Recognizer (SER) built on top of state-of-the-art IEMOCAP speech data-set which comprised of thousands of male and female speakers to derive the emotion of the participant throughout the sample. The final result procured are transformed MFCC features, and the probability of the sad emotion portrayed by the participant during the speech using the pre-trained model.

The pre-trained model used was based on the research conducted by the authors of

Chen et al. (2018), with an accuracy measure of 87.78%. The pre-trained model is a combination of Convolution Neural Network and Attention-model. This model takes MFCC features, Mel Spectrogram, and Zero-Cross energy to detect the emotion of a participant. The model is trained on the RAVDESS data-set, which was discussed in the Literature section. This data-set consists of equality in terms of Male and Female speakers to ensure a balanced data-set. The model was also tested on live-speech to ensure that the accuracy reported remains unbiased.

For the modelling of the data, all of the data were consolidated into an excel sheet where, first the heatmaps were augmented and normalized. A Resnet50 architecture-based model was used to process the flattened data, including the demographic features such as Age, Sex, self reported sadness levels before and after experiment, difference between sadness levels, and Memory test scores procured earlier, and Average distance from right eye and average distance from left eye procured from the dlib formula, and blink percentage, blink duration mean, blink duration standard deviation, saccade percentage, saccade duration mean, saccade duration , and standard deviation procured through the SMI software, and last but not the least, the 2 features from the speech sample, including the probability of the person showcasing a sad emotion, and MFCC features providing vital data on the speech sample of the respective participants.



(a) YOLOv5-based Screen Prediction (b) YOLOv5-based Circle Prediction

Figure 2: Examples of YOLOv5 Model outputs

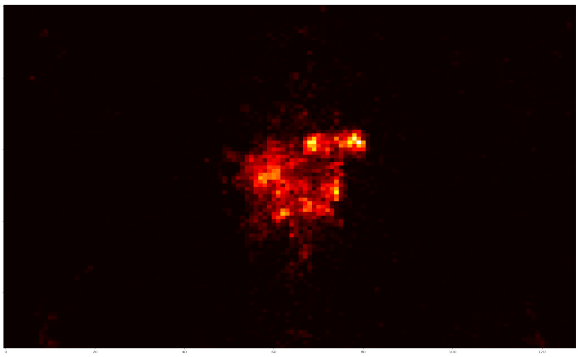


Figure 3: Heatmap

4 Design Specifications

The participants involved in the study were invited voluntarily.A pictorial representation of the process is as shown in Figure 4 Prior to the start of the experiments, each of these participants were asked to sign a consent-form stating that their participation is voluntary and that they consent to all the information that is being collected from them. They were also asked to provide demographic details such as their Age,Sex, and Sadness levels before the start of the experiment on a scale of 1 to 10. These features help understand and differentiate each person from another. Literature on the use of demographic information such as age and sadness prior to the experiment states that they can contribute significantly towards the prediction of empathy. The sadness levels were also an indicator of their sadness prior to the experiments, and their self-reported measure to indicate if there was any change in the same post watching the videos. The participants were then familiarized with the procedure and the working of the eye-tracking glasses. Post this, a 3-point calibration of the laptop screen was performed for each participant to ensure the information collected would be viable and reliable. The video stimuli was then shown to them, which was 13 minutes in duration comprising of 6 sad-emotion based stories narrated by actors directly looking at the camera. On completion, the participants were asked to update on their sadness levels after watching the video and subsequently asked to fill out a memory questionnaire with the following questions marked 1 mark, each, with no negative marking. This was done to ascertain their level of concentration during the experiment. The questions were :

- 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?

Each actor was provided with a fictional name to ensure that remembering the stories would be easier. Further, the participants were asked to answer one of the two following questions, for which their speech sample would be recorded in under one minute. The questions were:

- Could you relate to any video in particular of the ones shown to you? If So, why?

  OR

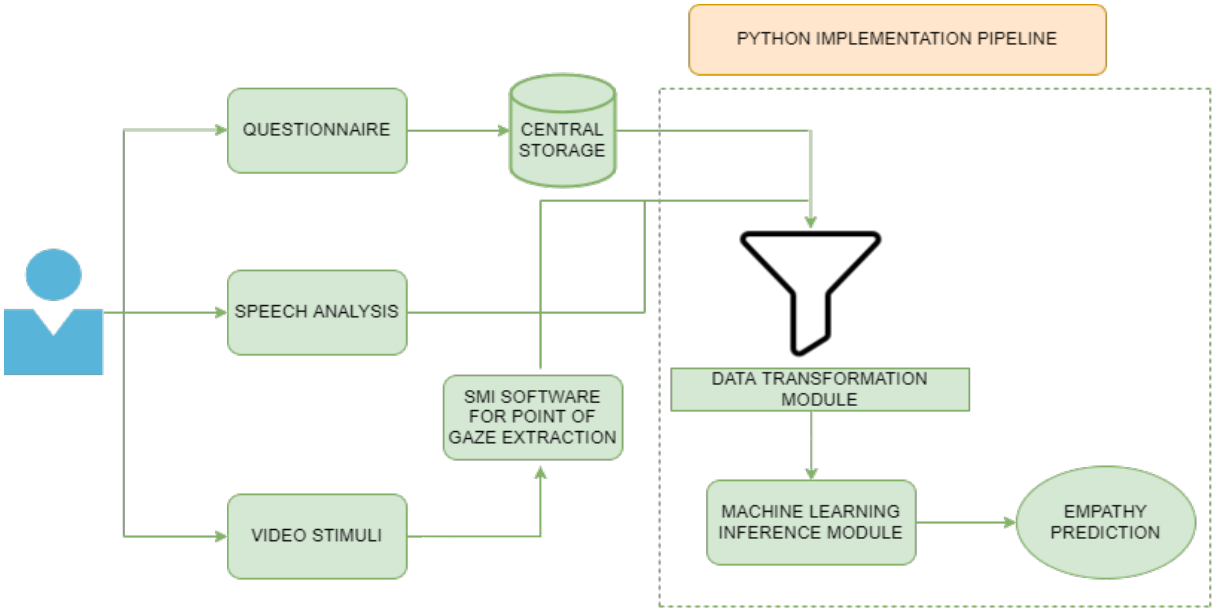


Figure 4: Research Methodology Work-Flow

- Would you be able to narrate a sad story from your life that you can think of at the moment?

The participants were given a choice to choose any one of these questions and answer accordingly. Their speech sample for the subsequent duration was recorded. On completion, the participants were asked to complete an empathy-based questionnaire which consisted of 10 questions, 5 related to Cognitive Empathy, and 5 related to Affective Empathy. The questionnaire was based on a 7-likert scale, where each question was presented with the following 7 choices, *"Disagree Strongly","Disagree Somewhat","Disagree Slightly","Neutral","Agree Slightly","Agree Somewhat","Agree Strongly"*. The marks associated with these questions were from the range of *-3* to *+3*, where the *-3* corresponded to lowest empathy level, and vice-versa. A score of "0" meant that the participant had a neutral emotion. Based on this, an empathy score for each participant was procured.

5 Implementation

This section comprises of the final model implemented to meet the research objectives. This is done by designing a machine learning pipeline which contains various features, some self-reported and some dependent on extraction from the SMI BeGaze software. A pictorial representation of the same is provided in Figure.5

The pipeline begins with some questions which the participant is asked to answer prior to the experiment. These features are the Age, Gender, Name, and their sadness levels prior to the experiments. The participant is then made to watch the videos, post-which the participant's speech sample is recorded, and then the empathy and memory questionnaires are answered. All of the data is then centrally stored in a data-base. Prior to storing the data on to the database, the details are collated based on the name of the participant. On completion, the name is replaced with a Unique Identifier Number. The name is then deleted for all records to maintain ethical considerations of the research.

The data acquired from the eye-tracking part of the process is then fed into the BeGaze

software, where the gaze pattern and the eye-tracking metrics is exported. Following this, the data from the database is combined to form a Python-based machine learning pipeline. There were two sadness levels recorded, prior to watching the videos are post watching the videos. A new feature calculating the difference of the same was obtained. The empathy-questionnaire data thus procured was also processed. The speech samples were then processed through the pre-trained model where the most dominant emotion of the participant throughout the narration is obtained. The speech sample was divided into frames od 25ms each. The emotion was recognized for each of these frames and finally collated based on count. The pre-trained model is capable of recognizing 6 basic emotions, *"Happy", "Sad", "Neutral", "Excited", "Disgusted", "Angry"*. The model classifies the frame emotion into one of these 6 emotions, and provides the result for the same. However, based on documentation obtained from the research, the emotions *"Sad", "Angry", and "Disgusted"* are combined together to provide the level of sadness in a participant. Based on this count, divided by the total number of frames, the values for the most dominant emotion of the participant for the entire duration of the video is procured. The eye-gaze data was then processed using the weights obtained from the YOLOv5 model which was trained to identify bounding boxes surrounding the point of gaze. The laptop screen data was essential in obtaining the heatmaps. The confidence interval for the bounding boxes and the laptop screen detection using the YOLOv5 models were 0.4 and 0.6 respectively. The Mean Average Precision (MAP) score for both of these models were greated than *0.995*. The training period for each of these models was 300 epochs, and the batch size was set to 8.

The test-data was then passed on to the machine learning model with the pre-trained weights producing the final result in the form of a binary classification of whether the participant is empathetic or not.

This entire pipeline is designed using Python as the coding language, and the empathy-questionnaire data is stored on OneDrive by Microsoft.

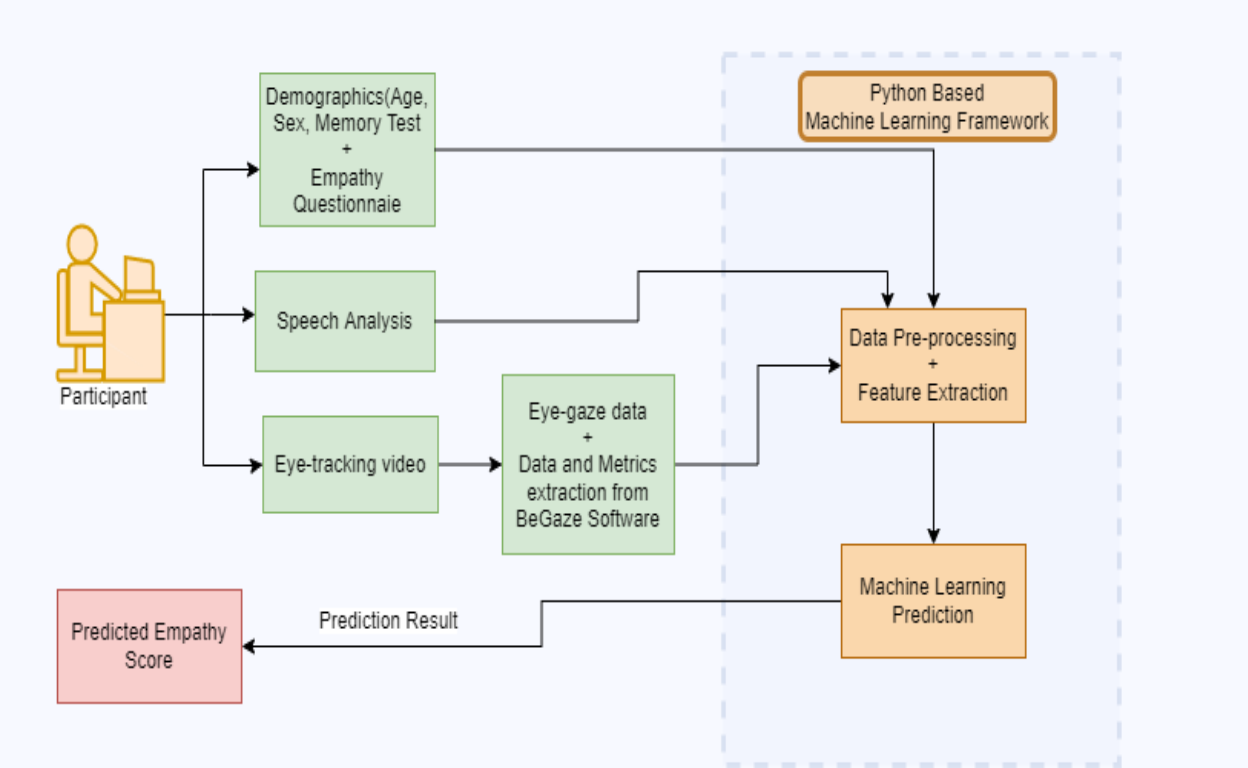


Figure 5: Implementation Overview

6 Evaluation

The total number of experiments conducted were 50, of which post data pre-processing and extraction, the total number of usable data-points were 44. This problem essentially occurred due to technical difficulties of the experiment with respect to the eye-tracking and storage issues with the phone. Hence, post the pre-processing phase, since parts of the data-point were corrupted, the features corresponding to all of these 6 participants was deleted prior to modelling.

The participants that had defined the Empathy-questionnaire data was collected and converted into a binary-class. The threshold was set at 1.3, which meant that all participants with a score greater than 1.3 were tagged as highly empathetic and the ones below that were tagged as non-empathetic. This resulted in a dataset of *18 highly empathetic people* and *26 non-empathetic people*. While, originally the questionnaire would consider them as low-to-moderately empathetic, the research objectives is to identify participants with high empathy levels. Hence, the 26 people were classified as non-empathetic or other. There were multiple performance metrics thus calculated for the various models were F1 score, Precision, and Accuracy. However, for the sake of this research, the F1 score was finalized as the metric of choice to determine the performance of the model. This was because F1 score combines and accounts for the precision and recall, whereas accuracy wouldn't account for them, and since the dataset is imbalanced with an approximately 40-60 split. The hyper-parameter tuning was set to random, and for all the models, the K-fold validation was set to 3.

The pictorial representations for the Memory Test Data, Figure 6, the levels of Sadness before, Figure 7a, and after watching the videos Figure 7b for all the participants is presented below.

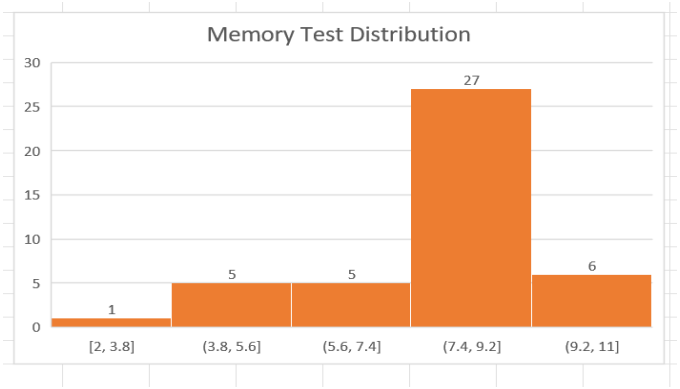
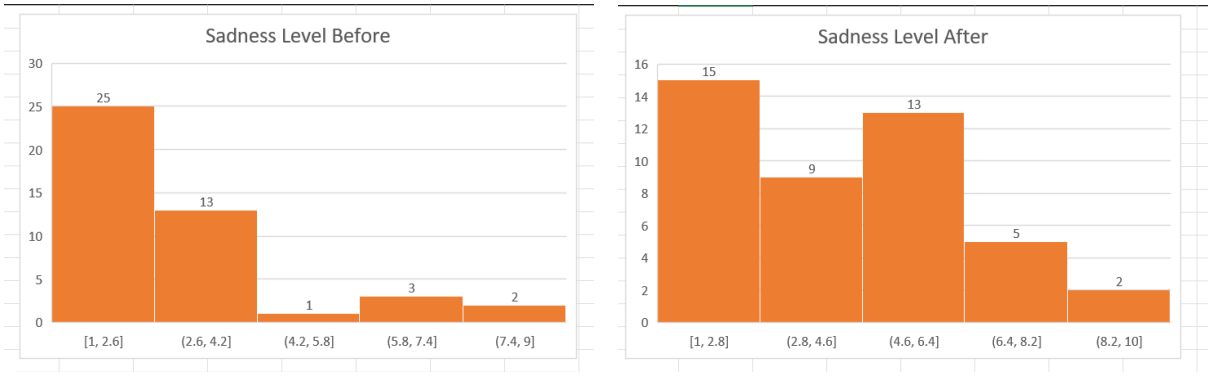


Figure 6: Memory Test Distribution



(a) Distribution of the Sadness Level Before Watching the Visual Stimuli

(b) Distribution of the Sadness Level After Watching the Visual Stimuli

Figure 7: Levels of Sadness Before and After

6.1 Experiment 1 - Logistic Regression with Demographic Features

The demographic details,sex, level of sadness before and after the experiment, difference in levels of sadness, and memory test scores that were collected from the participant was used to design and develop a Logistic regression-based classifier.

- C: This is a Regularization Parameter, which is responsible for over-fitting. The tuned value of C derived is - 0.02
- Solver: A Solver is responsible for choosing an algorithm that can resolve the optimization problem. The tuned value of the Solver is: newton-cg
- Penalty: This parameter is responsible for deciding the Penalty that would be applied on to Optimization problem. Penalty is: l2

The model achieved an accuracy of 72.2%, and F1 score of 0.21 on the test-data supplied.

6.2 Experiment 2 - Logistic Regression with Speech Samples

A logistic regression based classifier was trained on the feature procured from the pre-trained model on the speech samples. On completion, despite hyper-parameter tuning, the model under-performed with an accuracy measure of 43% and an F1 score of 0, indicating that the speech feature wasn't sufficient in determining the empathy of the participant, in the given context.

- C: This is a Regularization Parameter, which is responsible for over-fitting. The tuned value of C derived is - 0.08
- Solver: A Solver is responsible for choosing an algorithm that can resolve the optimization problem. The tuned value of the Solver is: newton-cg
- Penalty: This parameter is responsible for deciding the Penalty that would be applied on to Optimization problem. Penalty is: l2

6.3 Experiment 3 - Logistic Regression with features of Eye-tracking

A logistic regression-based classifier was trained on the features extracted using Begaze software, and the heatmaps generated using the YOLOv5 models. Furthermore, Principal Component Analysis was done to select 34 components that were shown to contribute a total of 95% of the variance. The data-set also consists of features such as blink percentage, saccade percentage and average distance from both eyes. The model was trained and tuned on these features.

- C: This is a Regularization Parameter, which is responsible for over-fitting. The tuned value of C derived is - 0.11
- Solver: A Solver is responsible for choosing an algorithm that can resolve the optimization problem. The tuned value of the Solver is: newton-cg
- Penalty: This parameter is responsible for deciding the Penalty that would be applied on to Optimization problem. Penalty is: l2

The accuracy obtained for this model was 59% with an F1 score of 0.58.

6.4 Experiment 4 - Logistic Regression with all data

For this final model, all of the above mentioned features was combined to train the model, namely Speech features, Eye-tracking features, and Demographic features. The model produced the highest accuracy of 89%, with an F1 score of 0.86, and recall of 0.75 and a precision of 1.

- C: This is a Regularization Parameter, which is responsible for over-fitting. The tuned value of C derived is - 0.17
- Solver: A Solver is responsible for choosing an algorithm that can resolve the optimization problem. The tuned value of the Solver is: newton-cg
- Penalty: This parameter is responsible for deciding the Penalty that would be applied on to Optimization problem. Penalty is: l2

6.5 Experiment 5 - Other Models on all data

Alongside Logistic Regression, Random Forest, Gradient Boosting and Artificial Neural Network-based model was deployed. The following are the results from them.

Logistic regression based model outperformed the rest with a precision score of 1.0, recall of 0.75 and an F1 score of 0.86. Gradient Boosting, a close second performed well with an F1 score of 0.57, precision of 0.67, a low recall of 0.5. Random Forest model had a high precision of 1.0, but a low recall and F1 score, which was 0.25 and 0.4, respectively. ANN performed the worst with an F1 score, precision and recall of 0.

The poor results from the Artificial Neural Networks were due to the lack of data for training. The number of data-points were too low for the model to learn anything. The model was created with 4 dense-layers, a Relu activation function, Softmax as the activation function for the last layer, and the loss function defined was binary cross-entropy since the final results were only 0 or 1.

The confusion matrices for the Logistic Regression (Figure 8), Random Forest (Figure 9), and Gradient Boosting (Figure 10) models are shown below. The confusion matrix for ANN is not displayed due to non-significance in the results obtained. The table collectively explains the results obtained from the different models for different features. Table 1

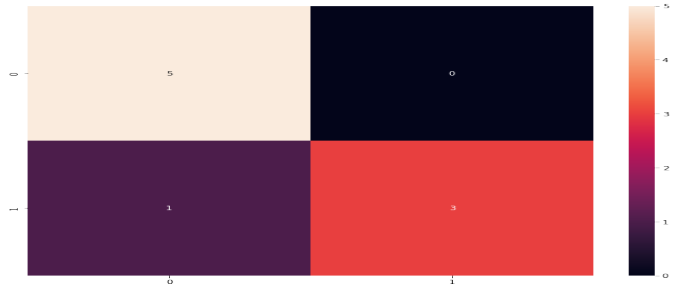


Figure 8: Logistic Regression Model

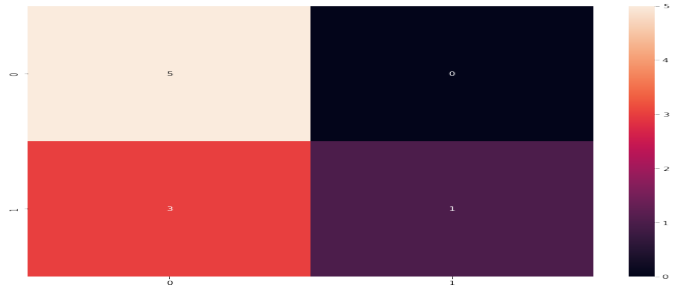


Figure 9: Random Forest Model

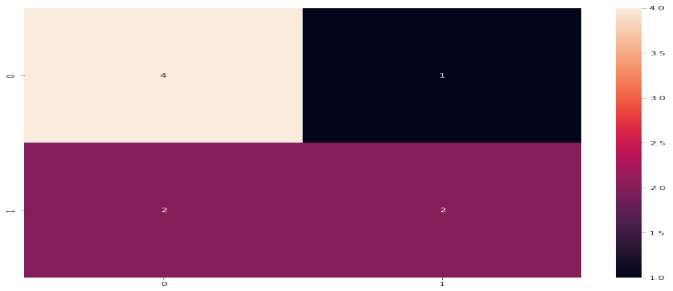


Figure 10: Gradient Boosting Model

Table 1: Comparative Study of Obtained Results

Features	ML Model	Accuracy	F1 score
Demographic Features	Logistic-regression	72.2%	0.21
Speech Signals	Logistic-regression	78.60%	0.28
Eye-tracking Features	Logistic-regression	59%	0.58
Combined Features	Logistic-regression	89%	0.87
Combined Features	Gradient-boosting	68%	0.57
Combined Features	Random-forest	51%	0.25

6.6 Discussion

The Logistic Regression model performed well with the features independently as well. However, it can be observed, of these, the eye-tracking features thus extracted contributed largely towards attaining a significant F1 score of 0.58, while the features extracted from the speech samples didn't seem to contribute much. The final model, thus supporting our initial hypothesis that the combination of all features would perform the best is proven true, with the best model being the Logistic Regression model with a combination of all features, with an F1 score of 0.87, an accuracy of 89%, precision and recall of 1.0 and 0.75, respectively.

The final framework consisted of Speech features extracted from the pre-trained model as well. However, the reason why they didn't work well independently was because there was only one feature that was extracted from the speech data. If other important features such as MFCC, Mel-Spectrograms, Zero-cross- Entropy were to have been extracted along with the Emotion recognized, the contribution would've been much larger and greater. The contribution of the memory test-scores along with other demographics were also non-significant as all the participants had a high score. However, the high score in itself tells us that all participants were concentrating and viewing the stimuli for the entire duration, thereby validating the experimental procedure thus imbibed.

The importance of all the features used in this research is portrayed in Fig.11. It is clear that the heatmaps, alongside the PCA features procured contributed highly towards obtaining the results. Furthermore, the contribution of the levels of sadness before and after, the average distance between the eyes, the correlation between empathy and eye-tracking thus hypothesized is true. While the speech signal's feature did not show high importance towards the results, it is still adds value to the overall model. The confusion matrices for the Logistic Regression, Random Forest, and Gradient Boosting models are

shown below. The confusion matrix for ANN is not displayed due to non-significance in the results obtained.

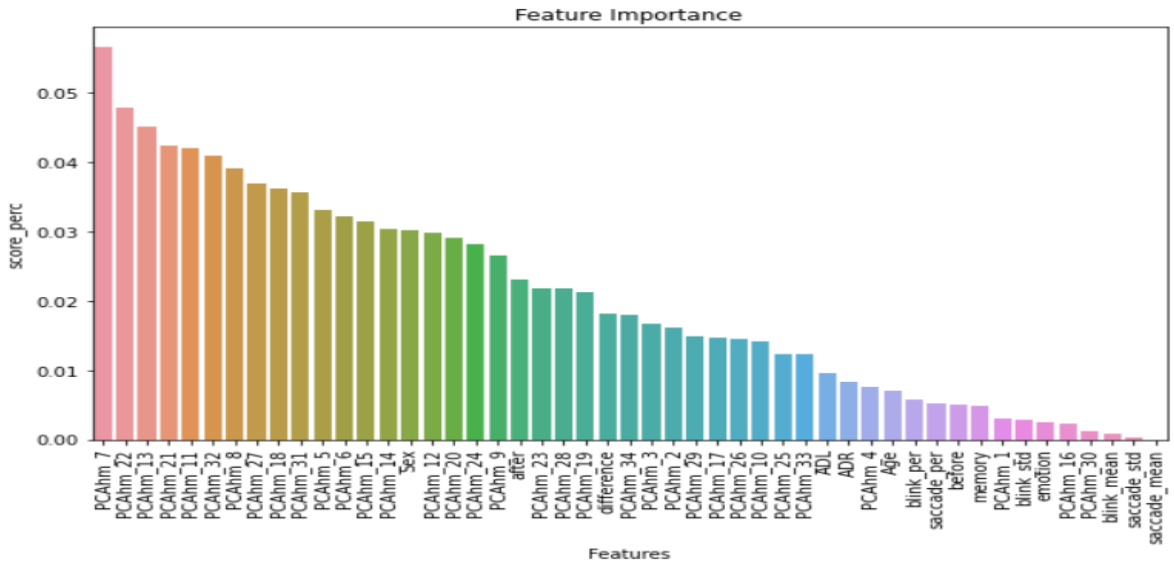


Figure 11: Importance of Features used in Modelling

7 Conclusion and Future Work

In conclusion, the research objectives, to design a machine learning framework to predict empathy with the help of Eye-tracking and Speech Analysis was achieved. The research study has provided significant insights that the point-of-gaze patterns is of significant value in determining the level of empathy in a person. Alongside this, the presence of features from Speech signals provides a significant value to the stack. The obtained results help us support the aforementioned arguments. Logistic-Regression based model provided the best results with an F1 score of 0.86 and an accuracy of 89%.

The research pipeline built can be bettered in three ways, first is the conduction of more experiments hence, collection of more data with varied stimuli.Secondly, the deployment of Convolution-based Neural Networks would provide significantly better results as CNNs are known to produce good result with images. Hence, the combination of heatmaps and speech signals-based Mel-spectrograms could be provided as an input to the CNN to procure better results. The use of CNNs could not be achieved in this research due to lack of significant amount of data. Lastly, the methodology used to extract heatmap information from the eye-tracked data could be bettered with the use of advanced algorithms and complex methodologies.

Furthermore, the deployment of a pipeline with the above mentioned additions, and more, would drastically increase and improve the quality of recruitment of medical professional where emphasis on prediction of Empathy is necessarily high.

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.

Casas, J., Spring, T., Daher, K., Mugellini, E., Khaled, O. A. and Cudré-Mauroux, P. (2021). *Enhancing Conversational Agents with Empathic Abilities*, Association for Computing Machinery, New York, NY, USA, p. 41–47.
URL: *https://doi.org/10.1145/3472306.3478344*

Chen, M., He, X., Yang, J. and Zhang, H. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, *IEEE Signal Processing Letters* **25**(10): 1440–1444.

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.

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.

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.

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.

Kammoun, M. and Ellouze, N. (2006). Spectral features detection of speech emotion and speaking styles recognition based on hmm classifier, *Proc. of WSEAS International Conference of Signal Processing*, Citeseer, pp. 27–29.

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.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H. and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review, *IEEE Access* **7**: 117327–117345.

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

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.

Morais, E., Hoory, R., Zhu, W., Gat, I., Damasceno, M. and Aronowitz, H. (2022). Speech emotion recognition using self-supervised features, *arXiv preprint arXiv:2202.03896* .

Nimmagadda, R., Arora, K. and Martin, M. V. (2022). Emotion recognition models for companion robots, *The Journal of Supercomputing* pp. 1–18.

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**.

Ooi, C. S., Seng, K. P., Ang, L.-M. and Chew, L. W. (2014). A new approach of audio emotion recognition, *Expert systems with applications* **41**(13): 5858–5869.

Přibil, J. and Přibilová, A. (2012). Formant features statistical analysis of male and female emotional speech in czech and slovak, *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, pp. 427–431.

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.

Schoneveld, L., Othmani, A. and Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition, *Pattern Recognition Letters* **146**: 1–7.

Singh, L. (2022). Deep bi-directional lstm network with cnn features for human emotion recognition in audio-video signals, *International Journal of Swarm Intelligence* **7**(1): 110–122.

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* .

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.