CLASSIFICATION OF ALZHEIMER'S USING DEEP-LEARNING METHODS ON WEBCAM-BASED GAZE DATA

by

Anuj Harisinghani

B.Tech., Amity University Uttar Pradesh, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2023

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Classification of Alzheimer's using Deep-Learning Methods on Webcam-based Gaze Data

submitted by Anuj Harisinghani in partial fulfilment of the requirements for

the degree of Master of Science

in Computer Science

Examining Committee:

Cristina Conati, Professor, Department of Computer Science, UBC
Supervisor

Thalia Shoshana Field, Associate Professor, Vancouver Stroke Program and Division of Neurology, Faculty of Medicine, UBC
Supervisory Committee Member

Abstract

There has been increasing interest in non-invasive predictors of Alzheimer's disease (AD) as an initial screening for this condition. Previously, successful attempts leveraged eye-tracking and language data generated during picture narration and reading tasks. These results were obtained with high-end, expensive eye-trackers. We explore classification using eye-tracking data collected with a webcam, where our classifiers are built using a deep-learning approach. Our results show that the webcam gaze classifier is not as good as the classifier based on high-end eye-tracking data, meaning its AU-ROC, Sensitivity and Specificity are significantly lower. However, the webcam-based classifier still beats a majority-class baseline classifier in terms of AU-ROC, indicating that predictive signals can be extracted from webcam gaze tracking. Our results provide an encouraging proof of concept that webcam gaze tracking should be further explored as an affordable alternative to high-end eye-trackers for the detection of AD.

Lay Summary

Dementia affects 55 million people globally according to the WHO. Alzheimer's Disease (AD) is the most common cause of dementia which generally affects older populations. Current screening tools for detecting AD are invasive and expensive, and so there exists the need for inexpensive, non-invasive and accurate screening tools. Recent developments in non-invasive methods involve high-end eye-trackers which have moderate accuracy in detecting AD (Jang et al., 2021), however, high-end eye-trackers are expensive and require in-person assessments. In this thesis, we explore the viability of webcam video recordings of clinically diagnosed patients and healthy participants as a low-cost alternative to high-end eye-trackers. We apply Deep Learning methods on these video recordings to show that although our solution using webcam video does not perform as well as high-end eye-trackers, it can extract meaningful predictive information and could potentially be improved to be an alternative to high-end eye-trackers.

Preface

This Master's thesis is the culmination of a research collaboration between the Department of Computer Science and Department of Medicine at UBC called the CANARY Cognition Group, who have collected the data described in Chapter 3. Implementation of the data pre-processing techniques, Machine Learning methods and statistical analysis were done by the author, Anuj Harisinghani. This work was supervised throughout by Professor Cristina Conati and was subject to constant feedback from the members of the CANARY Cognition Group. Anuj Harisinghani and Professor Cristina Conati wrote all parts of this work.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD – Alzheimer's Disease

AOI – Area of Interest

AU-ROC – Area Under the ROC Curve

BPTT – Back-Propagation Through Time

CV – Cross-validation

GRU – Gated Recurrent Unit

IReST – International Reading Speed Texts

LSTM – Long-Short Term Memory

ML – Machine Learning

MoCA – Montreal Cognitive Assessment

RF – Random Forest

RNN – Recurrent Neural Network

Acknowledgements

I would like to thank my supervisor Professor Cristina Conati for guiding me throughout the process of building the methods mentioned in this thesis and for helping me develop the mental tenacity needed for conducting research. I have grown a lot over the course of this program, thanks to the faculty of the Department of Computer Science, who have broadened my understanding of dozens of Computer Science concepts through the excellent courses they've taught.

I would also like to thank the amazing people at the Canary Cognition group, for providing me the data used in this thesis, as well as for the consistent feedback and support I received while working with them.

Finally, I want to thank my parents and my brother Garvit Harisinghani, who always believed in me and helped me get through the toughest phases of my life.

Chapter 1: Introduction

There has been increasing interest in devising lightweight predictors of Alzheimer's disease (AD) to be used as an initial screening for this condition. Assessments such as specialized neuroimaging and detailed cognitive assessments are invasive and require time, resources and expert personnel for administration. Existing brief cognitive screening tools (Cordell et al. 2013) solve the issue with invasiveness and intense time requirement, however, they have some limitations of their own. For example: the ceiling effect, where the tests may be insufficiently sensitive to detect mild cognitive issues in highly educated or high functioning individuals with early-stage AD.

Previous research has shown the potential of machine learning (ML) in predicting AD by leveraging eye-tracking alone, or together with language data generated during simple screening tasks (Biondi et al. 2018; Jang et al. 2021; Pavisic et al. 2017). However, these results were obtained with high-end, expensive eye-trackers that require patients to be evaluated in-person. There is a strong need for brief, sensitive automated cognitive tests that can be administered remotely and at low cost. Such tests would be suitable for screening individuals in research and clinical settings for cognitive dysfunction, and longitudinally monitoring neurological decline or improvement. Another advantage is the potential low-resource assessments that can be performed using the patients' own devices and without the need for expert personnel for administration. The ability to conduct remote testing is a cost-effective and efficient solution for health research and clinical care.

In this thesis, we take a step toward building these lightweight predictors of AD, by using eye-tracking data collected with a webcam. Using a webcam is cost-effective and can be done remotely, compared to eye-trackers used for collecting eye-tracking data in existing research on

AD predictors. For our investigation, we leverage a dataset collected by Jang et al. (2021), which includes both webcam recordings and high-end eye-tracking data. The data was collected while participants were performing simple tasks involving describing a picture and reading a short text. We train classifiers end-to-end using Deep Learning on gaze estimation vectors returned by OpenFace (Baltrusaitis et al. 2018), a standard software to extract relevant face properties from images. End-to-end training with low-level inputs is suitable due to the difficulty of defining informative high-level features from these data, which means we do not have the standard landmarks for attention tracking such as fixations and saccades. Since the sequences generated by OpenFace are extremely long, we apply a GRU model enhanced with a technique, target replication, which has been shown to better process long input sequences (Yue-Hei Ng et al. 2015). We compare our classifiers against the best-performing classifiers that were developed by Jang et al. (2021), by leveraging gaze data collected with a high-end Tobii X3-T120 during the same study. Our results show that a webcam gaze classifier does not perform as well as classifiers using high-end gaze data, however, it performs better than a majority class baseline. Our results provide an encouraging proof of concept for a webcam being an affordable alternative to high-end eye-trackers for prediction of AD. Our work alongside the work of (Hutt and D'Mello 2022), leverage gaze data from a webcam for user classification, suggesting that it is worthwhile to investigate if this easily available source of gaze data could be used for other classification tasks in which eye-tracking data has been shown to be highly beneficial. For example, Bixler and D'Mello (2016) detected mind-wandering in reading tasks using eye-gaze features. The study by Pusiol et al. (2016) analyzed the points-of-gaze of participants during social interactions to detect presence of Fragile X Syndrome (FXS), a genetic cause of autism. The works by Lallé, Conati, and Carenini (2016) and Sims and Conati (2020) predicted

occurrences of confusion when users interacted with the ValueChart interface, with the former using eye gaze features and the latter utilizing a neural network architecture.

Chapter 2: Related Work

Alzhiemer's disease (AD) has shown to affect functioning of the eye (Garbutt et al. 2008), which is observed through fixations, saccades, pupillary responses and other fundamental eye movement patterns. Supporting evidence can be found in the study by Molitor et al. (2015) wherein AD patients show abnormal saccadic behavior, saccadic intrusions and slowed pupillary responses in visual search and scene exploration tasks. In reading tasks, MacAskill and Anderson (2016) demonstrate that AD patients take longer to read text, fixate more, re-read words more frequently and are less likely to skip small words.

Based on this evidence, researchers have turned to eye movement analysis as a potential tool for classification of AD. Pavisic et al. (2017) trained hidden Markov models on a dataset of eye movement data from 36 individuals with young onset AD and 21 age matched controls and achieved 95% accuracy. Participants of the study completed a series of tasks, including smooth pursuit, saccadic and fixation stability tasks. (Biondi et al. 2018) report an accuracy of 87.78% using an autoencoder trained on a dataset that incorporates information on fixations, saccades and sentence length recorded while the participants completed a sentence reading task. The set of participants included 69 AD patients and 71 controls.

Jang et al. (2021) showed that combining eye movement and language features can successfully classify AD with 83.2% AU-ROC (more details on this work are provided in the next chapter, as we use the same dataset in this paper). All the existing work on the classification of AD from eye movements has relied on data captured with high-end eye-tracking equipment, whereas we want to ascertain if classification can be done with lightweight, low-cost webcam-based gaze tracking.

Gaze estimation from webcam images has been shown to be generally less accurate than when using high-end eye trackers (Kar and Corcoran 2017), although the difference in accuracy may decrease for tasks that require tracking attention to specific large areas of interest rather than to finer grain locations, such as attention to the camera (Zhang, Sugano, and Bulling 2017) or to a video conference window (Müller et al. 2018) or for extracting interpersonal gaze (Tran et al. 2022). There are also results on comparing gaze estimation from a smartphone camera with a Tobii Glasses Pro 2 wearable eye tracker, where Valliappan et al. (2020) trained a convolutional neural network (CNN) on the publicly available GazeCapture (Krafka et al. 2016) dataset, then fine-tuned it with user-specific calibrations (which led to a four-fold reduction in mean angular error). Participants completed a series of tasks on a smartphone: fixation tasks, which required participants to focus on a static visual target; visual search tasks, where participants were asked to locate a target object in an image; and reading comprehension tasks, where participants would read a given text and answer questions based on the text. The proposed CNN was able to achieve comparable accuracy (0.6-1 degrees error) to the Tobii Glasses Pro 2 eye tracker and the authors show that attention patterns generated by the smartphone-based eye tracking change with properties of the task (e.g., difficulty of the text to be read), suggesting that they could reproduce key findings from previous research using desktop eye trackers (such as detecting comprehension difficulty during reading).

The work of Hutt and D'Mello (2022) leverages higher-level fixation and saccade features generated using an unsupervised clustering algorithm from webcam-based gaze estimates to model user engagement and comprehension. On the other hand, there is substantial work on leveraging data on facial features from images captured via a webcam to classify a variety of viewer states and properties (e.g., disengagement from a learning task, (Boote,

Agarwal, and Mostow 2021)). However, in this work, gaze-related features are just a small subset of the many facial features generated by OpenFace, thus, unlike our work, it does not provide insights on to what extent web-based gaze data could be used as a substitute in classification tasks in which high-end eye-tracking data has been shown to be effective, such as for detecting mind-wandering (Bixler and D'Mello 2016), user-confusion (Lallé et al. 2016; Sims and Conati 2020) and classification of developmental disorders (Pusiol et al. 2016).

Chapter 3: Data Collection

In this paper we leverage data from a previous study (Jang et al. 2021) that evaluated the effectiveness of eye-tracking data collected with a high-end Tobii-Pro X3-120 eye-tracker as a predictor of AD. We will also use their best-performing Random Forest classifier as the gold standard to evaluate our classifiers. In this chapter, we summarize the aspects of the previous study that provide the context for our work.

Individuals with an existing diagnosis of AD or other mild cognitive impairments that are likely to develop into AD were recruited from a specialized memory clinic with a catchment of 4 million, while controls were recruited from the community, with efforts made for matching sex and age in the two groups. All participants were fluent in English, aged 50 or older and able to provide informed consent and to carry on a spontaneous conversation. The diagnoses for the clinic patients were made by expert clinicians with cognitive testing clinical data, and neuroimaging and laboratory data collected as per standard of care. All participants gave informed consent.

The cohort of participants used in this study includes 81 healthy controls (average age = 64.82, std. dev = 10.11) and 78 clinic patients (average age = 71.66, std. dev = 9.41). Participants were recruited between May 2019 and March 2020. For data collection, participants were seated at a testing platform, consisting of a monitor with a Logitech C922x ProStream webcam (attached on the top of the monitor) for recording the user's face and speech during the task and an infrared eye-tracker (Tobii-Pro X3-120) affixed at the bottom of the monitor to record gaze and pupil size data. Recordings for the webcam videos, audio files and raw eye-tracking files were started synchronously from the beginning of the assessment, and saved to the

experimenter's computer at the end of the assessment. Participants were asked to keep looking at the screen during the assessment and to avoid looking at the experimenter.

Participants performed four different tasks: a calibration task to ascertain the pupil size of the participant at rest; a picture description task; a reading task; a memory recall task that did not involve any visual component. Participants performed the tasks in order. The assessment took an average of 6 minutes to complete (std. dev = 2 minutes). In this paper, we leverage only the data collected during the picture description and the reading task (shown in Figure 1 (A) and (B)), because they are the ones involving extensive visual processing.

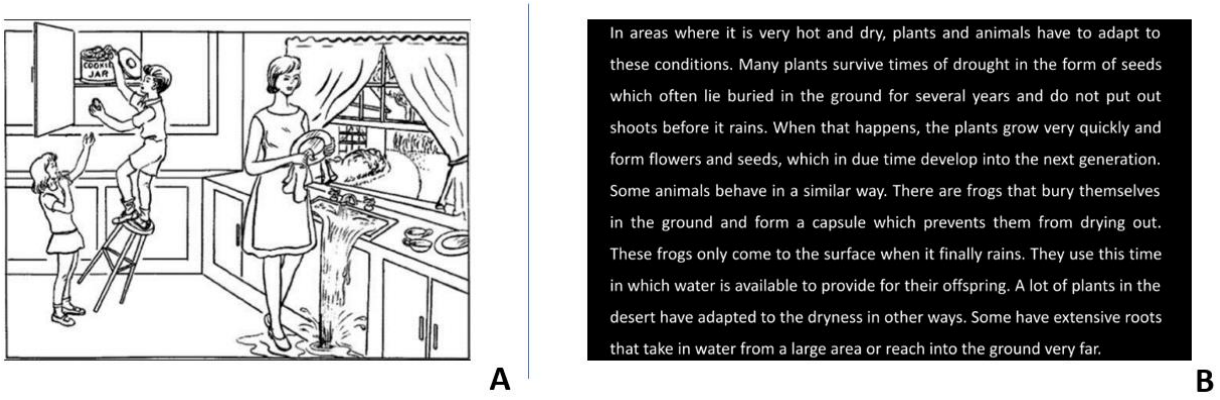

Figure 1: (A) Boston Cookie Theft picture for the picture description task; (B) Text shown for the reading task.

In the picture description task, participants describe the Boston Cookie Theft picture, a well-established speech task (Goodglass and Kaplan 1972). It is also a widely used and validated method for spontaneous speech assessment in a variety of clinical contexts, including Alzheimer's disease (Cummings 2019; Fraser, Meltzer, and Rudzicz 2016; Karlekar, Niu, and Bansal 2018; Kong et al. 2019).

In the reading task, participants read a standardized paragraph aloud from the International Reading Speed Texts (IReST), a collection of texts developed to be an assessment

tool for reading impairments, designed to be readable at a sixth-grade level (Trauzettel-Klosinski, Dietz, and the IReST Study Group 2012). The entire paragraph was presented to the participant all at once to recreate a natural reading task, similar to reading a newspaper or book. The goal of the reading task is to capture common reading-task deficits associated with AD, including reduced reading speed, and increased word fixations or re-fixations.

Chapter 4: Extracting Gaze Data Sequences from Webcam Videos

We used the open-source library OpenFace (Baltrusaitis et al. 2018) to capture gaze information from the videos of participants' faces. The videos were obtained concurrently with the eye-tracking and audio data as described in the previous chapter. Given a video frame of a face, OpenFace generates real-valued features like facial landmarks, head pose, eye gaze and facial action units. For the purpose of this paper, we focus on the gaze-related features. OpenFace captures gaze information in terms of two 3-dimensional vectors estimating the direction of the user gaze (x, y, and z coordinates) for each of the two eyes (See Figure 2). These data points are generated at 10Hz which is the sampling rate of the videos recorded by the webcam. We created a 9-valued vector from each sample, by concatenating the x,y,z coordinates for each eye, as well as the average of each coordinate between the two eyes.

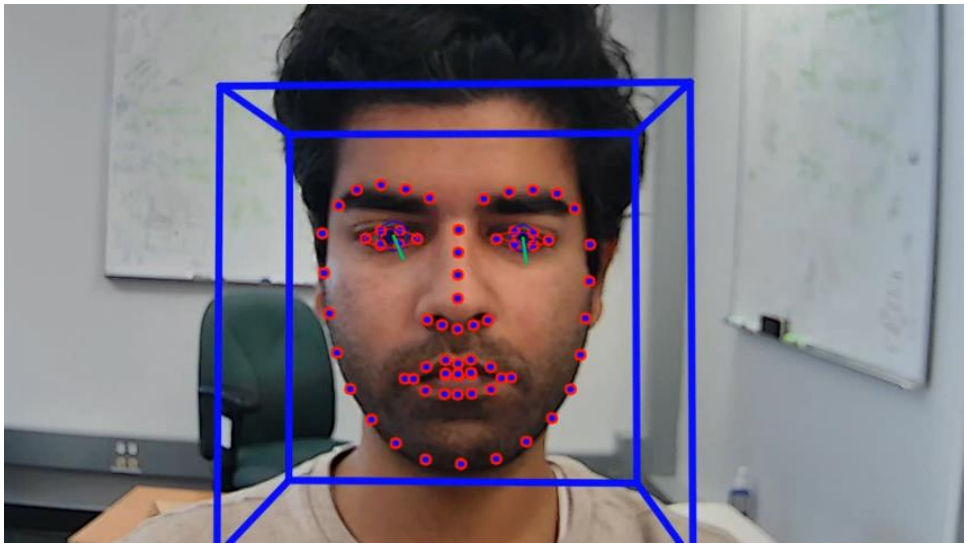

Figure 2: The eye-tracking setup. The image of the face is captured by the webcam and the gaze vectors from each eye are estimated by OpenFace as the ray originating from the pupil to the camera center. The person in the video is not a participant in the study due to privacy concerns.

Each of our processed sequences is for our two target tasks, Picture Description and Reading. Table 1 reports summary statistics on the length of these sequences. Figure 3 reports the distribution of these lengths over participants.

	Mean	Standard deviation	Median	Minimum	Maximum	Total	Number of Participants
Picture Description	750.94	529.56	608.5	130	3573	118,650	158
Reading	663.83	261.92	576	407	1971	105549	159

Table 1: Summary statistics of the length of sequences generated by OpenFace for each of the tasks

Our goal is to use these sequences in place of the eye-tracking sequences used in (Jang et al. 2021) to build classifiers of AD. The challenge is that we have much noisier data. Not only is this data sampled at 10Hz as opposed to the much higher sampling frequency of the Tobii eye-tracker (120Hz), but also the gaze data from OpenFace is naturally much less accurate with a mean angular error reported being 9.10 degrees (Baltrusaitis et al. 2018) while for Tobii Pro X3-120 it is less than 1 degree.

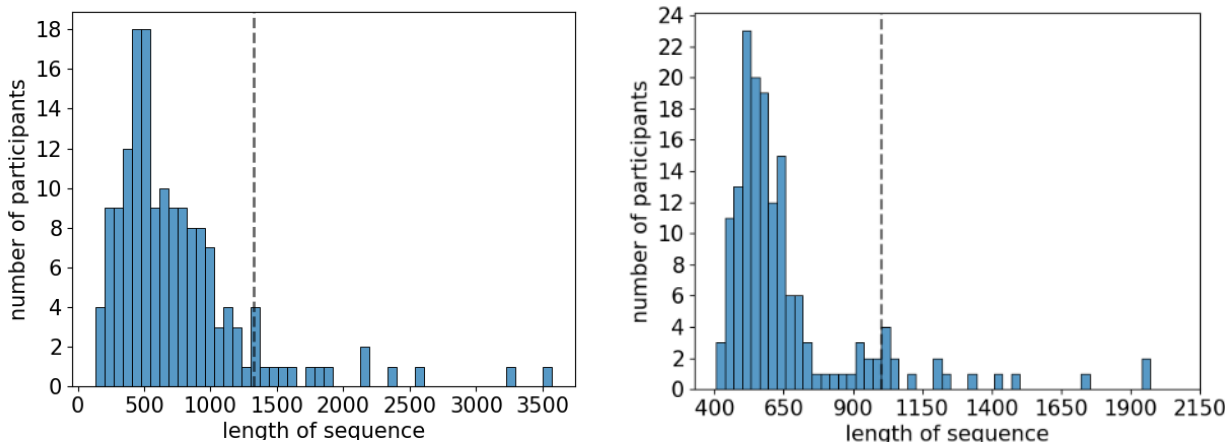

Figure 3: Histograms showing the distribution of sequence lengths over participants, for the Picture Description task (left) and Reading tasks (right). The dashed line marks the length of the 90-percentile participant for each task.

Chapter 5: Our Proposed Machine Learning Approach

To summarize, our goal is to use the gaze data derived via OpenFace from the videos of the study participants' faces (which we will refer to as the webcam dataset from now on) to build lightweight classifiers of AD, and to see if these lightweight classifiers can approximate the performance of the best classifiers obtained by Jang et al. (2021) when leveraging high-end eye-tracking data (which we will refer to as the Tobii dataset from now on)[1] . Their best performing classifiers were built using a non-Deep Learning machine learning algorithm, Random Forest, and engineered features (i.e., summary statistics) derived from the eye-tracking measures returned by the Tobii eye-tracker, such as fixation position and duration, saccades lengths etc. The features also relied on having information on relevant Areas of Interest (AOI) in the visual stimuli. Examples of these AOI-based features include the percentage of fixations detected in each of the AOIs, the longest fixation on the AOI, the time before the first fixation on each AOI, etc.

With the webcam dataset, we do not have information on meaningful attention-related constructs such as fixations and saccades, nor do we have a reliable way to capture which AOIs the user is looking at. All we have is the sequence of 9-dimensional vectors of coordinates derived from OpenFace, as we described above. Therefore, we choose not to try and define meaningful engineered features from these vectors and instead we leverage deep-learning to process them end-to-end using low-level inputs.

---

[1] This classifier which used the eye-tracking dataset from Tobii was trained using 126 of the 162 available participants, due to data validity issues in the eye-tracking data of the remaining participants. The webcam dataset, however, uses 159 of the 162 available participants, due to issues with the reported start and end times of the tasks for the 3 remaining participants. 1 participant was considered an outlier only in the Picture Description task due to their extremely long sequence, leaving 158 and 159 participants for Picture Description and Reading tasks, respectively, as seen in Table 1.

We look at Recurrent Neural Networks (RNNs) as they are suited for processing sequences of temporal data. Specifically, we use Gated Recurrent Unit (GRU) (Cho et al. 2014) networks, which are better than simple RNNs at dealing with longer sequences but are computationally more efficient than Long Short-Term Memory (LSTM) for a dataset with the same size as ours (which is small for Deep Learning standards) (see Table 1 and Figure 3). Figure 4 shows an example of sequence classification with a GRU: a sequence with n real-valued vectors x_i is fed one step at a time to the GRU. At the last timestep, the network makes a prediction $\hat{y}_n$ on the class of the sequence. Loss is calculated by comparing the prediction at the final timestep to the actual class, and Back-Propagation Through Time (BPTT) updates the network's hidden states using the loss.

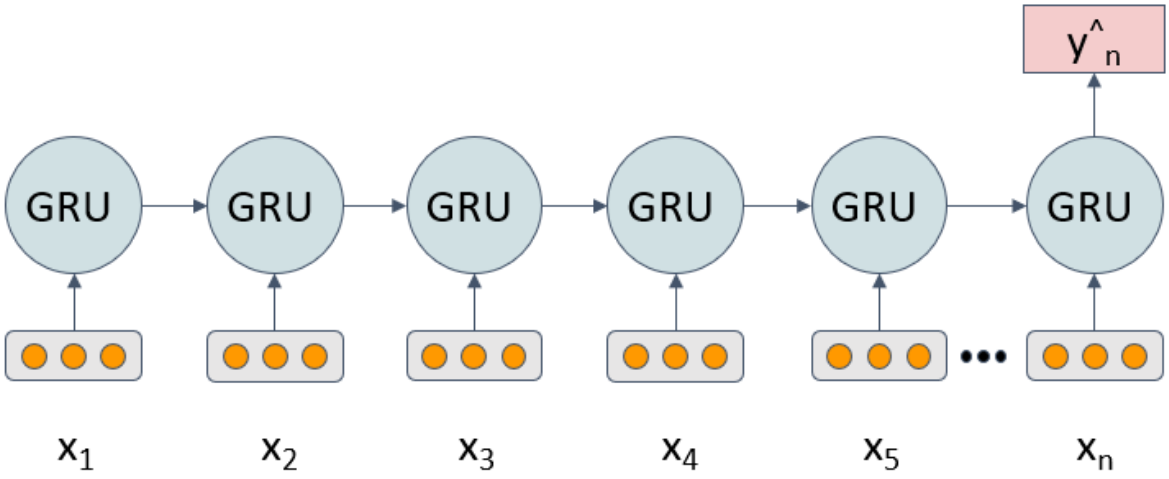

Figure 4: Vanilla sequence classification using a GRU. Each input is a real-valued vector (denoted by x_i) and the final prediction is $\hat{y}_n$.

Despite GRUs being good at dealing with long sequences, they still struggle to process sequences longer than 1,000 elements (Li et al. 2018), and even within this length, their

performance may negatively be impacted as the length increases. We address this issue in two ways: first, we shorten the sequences as much as possible; next, we augment the GRU with a mechanism to deal with long sequences (as discussed in the section below). Transformers, which are known to be better than RNN-based models on many tasks involving sequences, were also considered. However, to our knowledge, there is no pretrained transformer model for gaze data from a webcam, and our dataset is too small for learning parameters in a transformer model from scratch.

5.1 Reducing sequence length

Our goal here is to reduce the length of our sequences to be closer to the limit of 1,000, without impacting too much the amount of gaze information that we have for participants. Because sequence lengths in our dataset have a long-tailed distribution (see Figure 3), we identified participants who had a sequence length greater than the participant at the 90-percentile mark (1,323 for the Picture Description task and 1,001 for the Reading task), After finding these participants (16 in each task) we truncated their sequences to the aforementioned 90% percentile values instead of removing them outright, in order to retain as much data as possible. We perform the truncation at the end of the outliers' sequences rather than at the beginning to capture the important stage of setting up the visual scene which all participants must go through. The choice of the 90-percentile value was made because it strikes a balance between bringing the sequence lengths closer to the recommended sequence length limit of 1,000 and not losing potentially useful information in the gaze traces by truncating them too much.

Figure 5 and Table 2 show the updated summary statistics and distributions after truncating all sequences to the target 90-percentile length. The spike in the number of

participants in Figure 5 denotes all the sequences longer than the 90-percentile value being truncated to match the 90-percentile target.

	Mean	Standard deviation	Median	Minimum	Maximum	Total	Number of participants
Picture Description	684.80	336.45	608.5	130	1,323	108,199	158
Reading	633.50	167.33	576	407	1,001	100,728	159

Table 2: Updated summary statistics after truncation at the end for all sequences to the 90-percentile mark

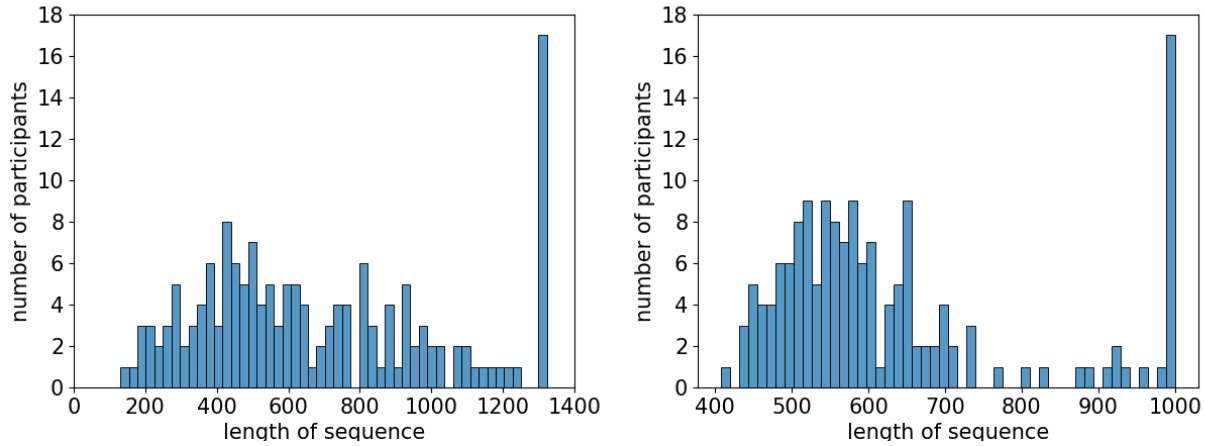

Figure 5: Histogram showing the distribution of sequence lengths over participants after truncation at the 90-percentile mark, for the Picture Description task (left) and Reading task (right).

Figure 6 shows the distribution of removed segments over the 16 outliers, for each task, expressed in terms of the percentage removed from the original sequence. We notice that in the Picture Description task, there are 2 participants who have about 60% of their sequences removed and 3 participants who have between 40-50% of their sequences removed. In the Reading task, 3 participants have between 40% and 50% of their sequences removed. However, even if we have lost large percentages of data for these participants, we still retain 1,323 data

points for Picture Description (i.e., about 132 seconds worth of interaction data) and 1,001 for Reading (i.e., about 100 seconds worth of interaction data) which we hope will be sufficient to capture interesting patterns.

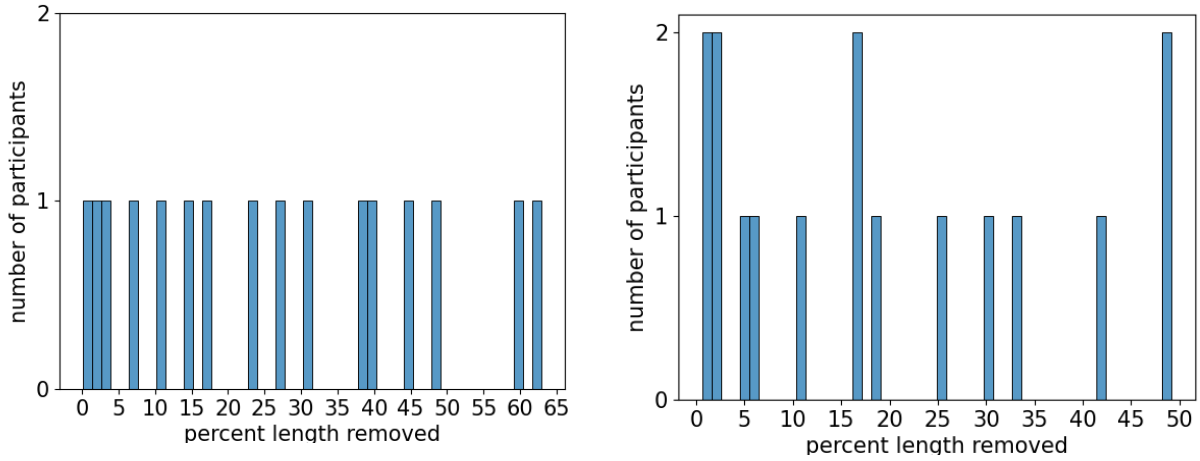

Figure 6: Histograms showing the distribution of percentages of discarded data in sequences due to truncation, over participants.

5.2 Augmenting the GRU with Target Replication

Even after the pre-processing steps described in the previous section, our sequences are over 600 steps on average, with peaks between 1,000 and 1,300 between the two tasks. Thus, we augment our GRU model with a mechanism that aims to alleviate issues that might arise due to excessive sequence length. This mechanism was first introduced by Yue-Hei Ng et al. (2015) to perform video classification with full videos up to 300 frames long, as opposed to the short video clips that were considered in the literature until then. The mechanism involves calculating and backpropagating the loss with respect to the target label at every timestep instead of only at the end of the sequence (see Figure 7 (A)). Because these intermediate backpropagations update the network parameters as the GRU sequentially processes the input, they help retain the information

from earlier timesteps which might otherwise get lost if backpropagation is performed only once from the end of a long sequence (Yue-Hei Ng et al. 2015). This approach has also been applied to process sequences of multivariate medical data (Futoma et al. 2017; Lipton et al. 2015), webcam video sequences for user disengagement (Boote et al. 2021), video sequences for multi-label object detection (Tripathi et al. 2016), and unstructured textual notes for patient mortality prediction (Grnarova et al. 2016).

All this previous work performed backpropagation of the loss at every timestep, which we will call Target Replication following Lipton et al. (2015). However, when we tried to apply this approach to our dataset, we found that replicating targets at every step is computationally very expensive, meaning that doing training and testing with cross-validation over our Webcam dataset would take more than 10 days when run on our available hardware consisting of four NVIDIA RTX A5000 GPUs. Therefore, we experimented with replicating targets at every k timestep (as shown in Figure 7 (B)), trying with values of k=50, 100, and 200. We found minimal differences in performance; thus, we settled on k=100 as a proof-of-concept for training the models described in the rest of this section (using this value reduced the computation time of cross-validation to less than a day). Detailed information about the implementations can be found in **Error! Reference source not found.**.

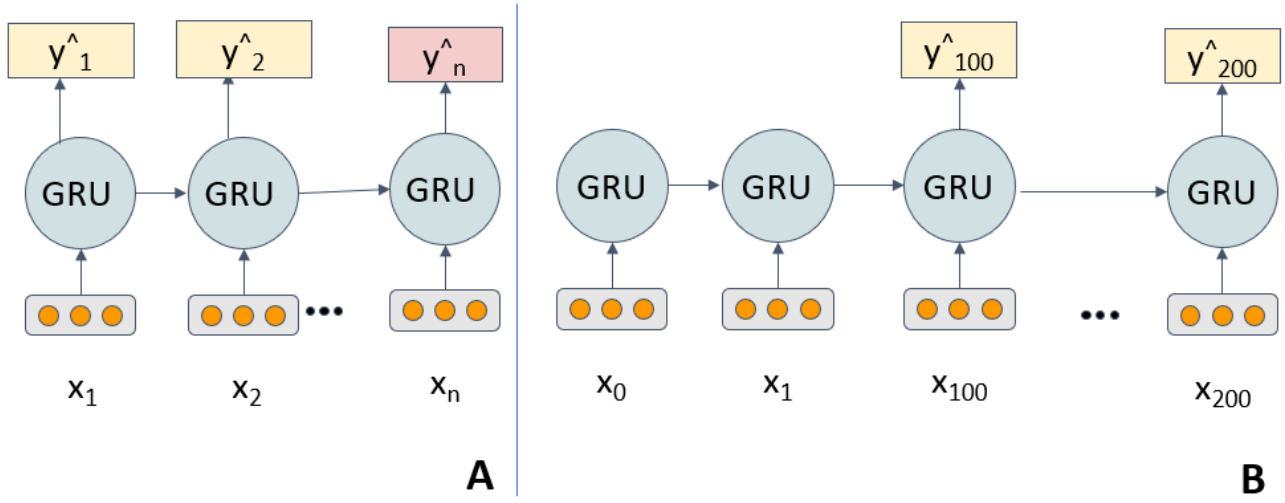

Figure 7: (A) Sequence classification using a GRU with Target Replication at every step. (B) Variation with Target Replication at every 100 timesteps.

Chapter 6: Results

In this chapter we present the classification performance achieved by our model (denoted by Webcam-GRU) on each of the two study tasks (Picture Description and Reading). We compare these results against the classification performance of the best-performing classifier from (Jang et al. 2021), namely a Random Forest classifier trained on data from the Tobii eye-tracker in (Jang et al. 2021) (denoted by Tobii-RF), and against a baseline classifier (denoted by Baseline) which always predicts the majority class on the same dataset used by the Webcam-GRU classifier.

We evaluate all our Webcam-GRU models using 3 runs of 10-fold cross-validation on our dataset of 159 participants. The relevant network parameters (such as number of layers, number of hidden cells, dropout rate) were selected in the nested cycle of each fold. Network outputs were made to 2 nodes with SoftMax activation. The Adam optimizer was used, and each network was trained for 500 epochs with learning rate set to 0.001. On the other hand, Tobii-RF and Baseline were trained using a 10-run 10-fold cross-validation procedure.

We use the following metrics to describe model performance:

1. Area Under the ROC Curve (AU-ROC) as a measure of the overall accuracy of each classifier in distinguishing between patients and healthy controls
2. Sensitivity (or true positive rate) - the proportion of patients that are correctly identified as such. It indicates the ability of the model to detect patients.
3. Specificity (or true negative rate) - the proportion of healthy controls that are correctly identified as such. It indicates the ability of the model to avoid false positives.

All the classifiers we test output a probability of a given participant being a patient. Sensitivity and specificity are based on using 0.5 as the probability threshold to classify a participant being a patient (positive class), as done in (Jang et al. 2021). Statistical comparisons

are made for AU-ROC, Sensitivity and Specificity metrics between the Webcam-GRU model and the Tobii-RF; and between the Webcam-GRU model and the Baseline[2], for each task separately. Comparisons between Tobii-RF and Baseline have established that Tobii-RF is statistically better than the baseline, as done in (Barral et al. 2020), hence, they were not included in this thesis.

We used the Mann-Whitney U test because, for all three target measures, the data were not normally distributed, making a standard t-test unsuitable for the statistical analysis. We applied a Benjamini-Hochberg adjustment to account for family-wise errors due to looking at three different metrics and two pairwise comparisons. We set $\boldsymbol{\alpha}$=0.05 and report the U statistic (U), adjusted p-values (p) and Cohen's d effect sizes (d).

6.1 Performance of the Webcam-GRU classifier

6.1.1 Results for the Picture Description (PD) task

Table 3 and Figure 8 show the performance of Webcam-GRU, Tobii-RF and Baseline for the eye modality of the Picture Description task. The pairwise comparison between the Webcam-GRU and Tobii-RF shows that Tobii-RF performs significantly better than Webcam-GRU in all metrics: AU-ROC ($U = 0$, $p < 0.01$, $d = -9.2$), Sensitivity ($U = 4$, $p < 0.05$, $d = -2.32$) and Specificity ($U = 0$, $p < 0.01$, $d = -3.43$). Pairwise comparison of Webcam-GRU against the Baseline shows a significant difference in all three metrics, with Webcam-GRU being better in

[2] It is important to note that the Baseline predicts "healthy control" for every data point, since our dataset of 159 participants consists of 81 healthy controls and 78 patients. Hence, only the AU-ROC metric is considered for faithfully comparing Webcam-GRU against the Baseline, since the Baseline will always score 0% for Specificity and 100% for Specificity due to its nature of always predicting the majority class.

AU-ROC (U = 0, p < 0.001, d = 3.68) and Sensitivity (U = 0, p < 0.001, d = 10.24), however, the Baseline has a significantly higher Specificity (U = 0, p < 0.001, d = -14.38).

	AU-ROC	Sensitivity	Specificity
Webcam-GRU	0.55±0.03	0.49±0.11	0.62±0.06
Tobii-RF	**0.77±0.02**	**0.62±0.04**	0.76±0.03
Baseline	0.50±0.00	0.0±0.00	**1.00±0.00**

Table 3: Performance results for Webcam-GRU, Tobii-RF and Baseline in the eye modality for the Picture Description task. Values are in terms of metric_mean(±sd). Bold entries denote the highest-performing classifier.

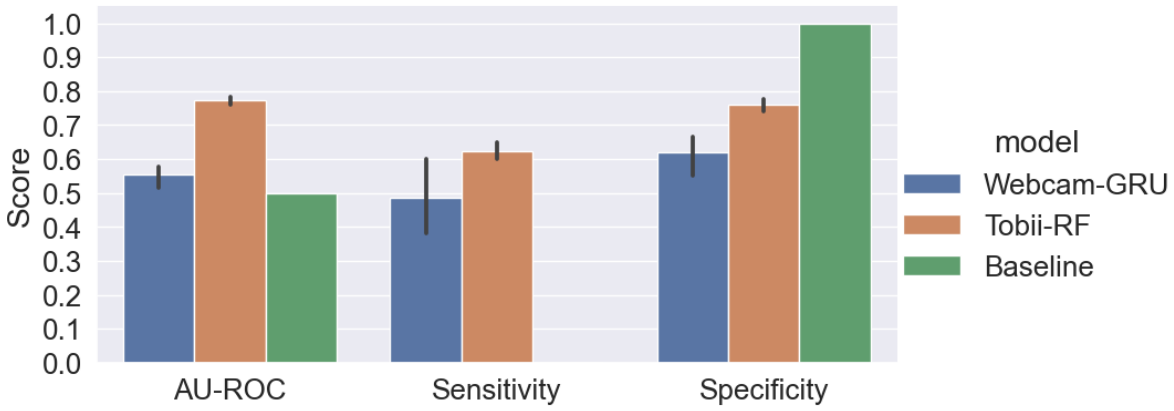

Figure 8: Performance of Webcam-GRU, Tobii-RF and Baseline in the eye modality Picture Description task in AU-ROC, Sensitivity, and Specificity.

6.1.2 Results for the Reading task

Similar to the findings from the Picture Description task, the pairwise comparison between the Webcam-GRU and Tobii-RF for the Reading task (Table 4, Figure 9) shows Tobii-RF performing significantly better than the Webcam-GRU in all metrics: AU-ROC (U = 0, p < 0.01, d = -6.01), Sensitivity (U = 3, p < 0.05, d = -1.63) and Specificity (U = 0, p < 0.01, d = -2.91. Comparison between the Webcam-GRU and Baseline show statistical difference in all three

metrics, with Webcam-GRU performing better in AU-ROC (U = 0, p < 0.001, d = 14.15) and Sensitivity (U = 0, p < 0.001, d = 54.88). The Webcam-GRU, however, performs significantly worse in Specificity (U = 5, p < 0.001, d = -27.04).

	AU-ROC	Sensitivity	Specificity
Webcam-GRU	0.59±0.01	0.54±0.02	0.63±0.03
Tobii-RF	**0.74±0.03**	**0.61±0.05**	0.71±0.03
Baseline	0.50±0.00	0.0±0.00	**1.00±0.00**

Table 4: Performance results for Webcam-GRU, Tobii-RF and Baseline in the eye modality for the Reading task. Values are in terms of metric_mean(±sd). Bold entries denote the highest-performing model.

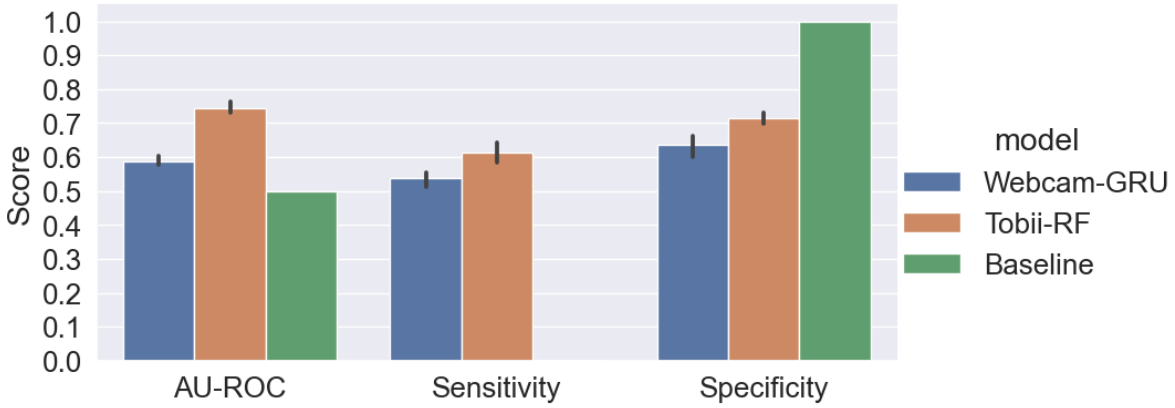

Figure 9: Performance of Webcam-GRU, Tobii-RF and Baseline in the eye modality Reading task in AU-ROC, Sensitivity and Specificity.

6.1.3 Discussion

Our results highlight an interesting difference in the performance of the Webcam-GRU in the two tasks, with a better outcome for the Reading task. A reason for this result could be that in (Jang et al. 2021), the most predictive features for the Tobii-RF classifier in the PD task were reported to be some of those related to Areas of Interest (AOI) whereas this was not the case for

the Reading task. Therefore, the fact that the Webcam-GRU model has no information about AOIs hampers its performance in the PD task but not in the Reading task. As for the reason why AOI information is more important for classification in the PD task than in the Reading task, this could be due to the fact that the Reading task is a smooth-pursuit gaze task in which all participants have to read the same standard text, which leads to similar gaze patterns across participants. On the other hand, the PD task is much more open-ended, with many different ways to scan the image to describe it, which leads to more varied gaze movements across participants. Information on AOIs may provide a way to capture regularities in these more varied patterns.

6.2 Classification performance for an ensemble of eye and language data

In this section we compare the performance of the Webcam-GRU classifier and the Tobii-RF classifier when each is used in combination with the language-based classifiers presented in (Jang et al. 2021) to create an ensemble predictor for AD. The gaze-based and language-based models are combined using the same approach as in (Jang et al. 2021): for each participant with data in both modalities (159), the prediction probabilities generated by the gaze-based and language-based classifiers are averaged. For the participants who only have valid data in one of the two modalities (0 for gaze and 3 for language), we take the prediction generated by the respective relevant classifiers. We will denote the language-based classifier as Lang from now on. Similar to the eye-only classification in section 6.1, Webcam-GRU is also compared against a Baseline which predicts the majority class.

6.2.1 Results for the Picture Description (PD) task

Statistical comparisons between Webcam-GRU + Lang and Tobii-RF + Lang in the PD task show that Tobii-RF + Lang is better in all three metrics, with a significant difference in AU-ROC

(U = 0, p < 0.001, d = -12), Sensitivity (U = 0, p < 0.001, d = -4) and Specificity (U = 0, p < 0.05, d = -2.92). Comparison between Webcam-GRU + Lang and Baseline show a similar result to the eye modality results of the PD task: AU-ROC (U = 0, p < 0.001, d = 13.59) and Sensitivity (U = 0, p < 0.001, d= 15.76) are significantly better for the Webcam-GRU + Lang model, but Specificity (U = 0, p < 0.001, d = -16.52) is significantly worse than the Baseline. Table 5 and Figure 10 show the performance of the combined models.

Even if Webcam-GRU still does not perform as well as Tobii-RF in detecting AD patients, we observe some improvement after combining Webcam-GRU with the language-based classifier. While the improvements from Webcam-GRU to Webcam-GRU + Lang are not statistically significant, the effect size for AU-ROC was large (d=1.04) and medium and small for Sensitivity (d=0.47) and Specificity (0.16) respectively. On the other hand, Tobii-RF + Lang shows significant improvement over Tobii-RF, which is consistent with the finding from (Barral et al. 2020).

	AU-ROC	Sensitivity	Specificity
Webcam-GRU + Lang	0.58±0.01	0.53±0.07	0.63±0.05
Tobii-RF + Lang	**0.78±0.02**	**0.69±0.03**	0.74±0.03
Baseline	0.50±0.00	0.00±0.00	**1.00±0.00**

Table 5: Performance results for Webcam-GRU + Lang., Tobii-RF + Lang. and Baseline for the Picture Description task. Values are in terms of metric_mean(±sd). Bold entries denote the highest-performing model.

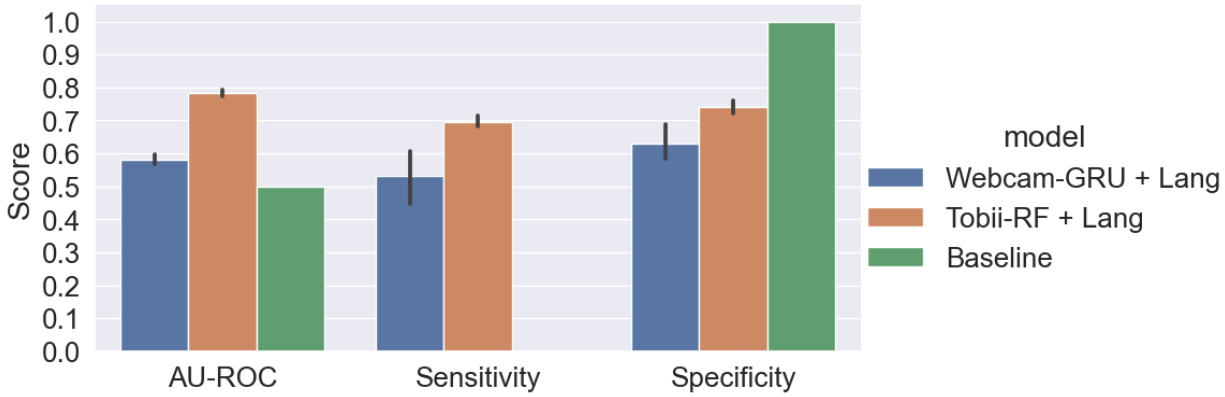

Figure 10: Performance of Webcam-GRU + Lang., Tobii-RF + Lang. and Baseline in the Picture Description task in AU-ROC, Sensitivity and Specificity

6.2.2 Results for the Reading task

Statistical comparisons between Webcam-GRU + Lang and Tobii-RF + Lang in the Reading task reveal the same characteristics as the eye modality comparisons: the model with Webcam-GRU is worse in all metrics: it is significantly worse in AU-ROC ($U = 0$, $p < 0.01$, $d = -9.59$), Sensitivity ($U = 0$, $p < 0.01$, $d = -3.15$) and Specificity ($U = 0$, $p < 0.01$, $d = -3.11$). Comparison between Webcam-GRU + Lang and the Baseline show significant difference in all three metrics, with Webcam-GRU + Lang performing better in AU-ROC ($U = 0$, $p < 0.001$, $d = 11.24$) and Sensitivity ($U = 0$, $p < 0.001$, $d = 27.86$), but performing worse in Specificity ($U = 0$, $p < 0.001$, $d = -16.21$). Table 6 and Figure 11 show the performance of all models mentioned above.

Similar to the Picture Description task, combining Tobii-RF with Language significantly improves performance. On the other hand, combining language to Webcam-GRU doesn't show statistically significant improvement over the non-language Webcam-GRU, however, the effect size is very large for AU-ROC ($d=1.62$) and Sensitivity ($d=1.3$), and medium for Specificity ($d=0.38$).

	AU-ROC	Sensitivity	Specificity
Webcam-GRU + Lang	0.62±0.02	0.58±0.04	0.65±0.05
Tobii-RF + Lang	**0.80±0.02**	**0.67±0.02**	0.79±0.04
Baseline	0.50±0.00	0.00±0.00	**1.00±0.00**

Table 6: Performance results for Webcam-GRU + Lang., Tobii-RF + Lang. and Baseline for the Reading task. Values are in terms of metric_mean(±sd). Bold entries denote the highest performing model.

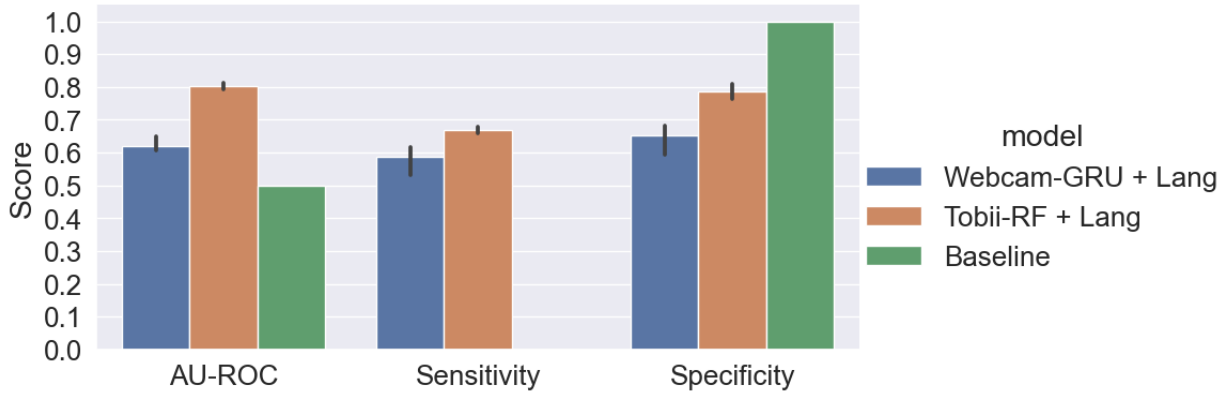

Figure 11: Performance of Webcam-GRU + Lang., Tobii-RF + Lang. and Baseline in the Reading task in AU-ROC, Sensitivity and Specificity.

In a nutshell, the results for the ensemble of eye and language data mirror the ones leveraging only eye data, with the addition of language benefitting only the Tobii-RF classifier.

Chapter 7: Conclusions and Future Work

In this thesis, we presented a classification model that uses webcam-based eye-tracking data for the classification of Alzheimer's disease (AD). Our goal is to contribute to the development of lightweight, low-cost predictors that can be used as an initial screening for this condition. We build on existing work which successfully leveraged gaze data obtained with high-end, expensive eye-trackers for AD classification, where the data was collected during simple picture description and reading tasks. Our results show that our webcam gaze classifier (Webcam-GRU) does not perform as well as the classifiers using high-end gaze data in both Picture Description and Reading tasks. In contrast, the webcam gaze classifier is significantly better than a majority class Baseline, showing that Webcam-GRU is able to find predictive signals from webcam-based gaze data. Our results provide encouraging novel evidence that using a common and affordable webcam instead of an expensive high-end eye-tracker may be a future viable solution to develop more accessible screening tools for dementia.

Our results contribute insights on the potential of web-based gaze data as a substitute in user classification tasks for which high-end eye-tracking data was shown to be effective (e.g., detecting mind-wandering, user-confusion developmental disorders and user perceptual abilities). In particular, we experimented with using deep learning models to train our classifiers end-to-end. This choice was made to circumvent the difficulty of identifying meaningful features for training, given that our web-based gaze data does not provide information on constructs such as fixations, saccades and Areas of Interests, which are commonly used to define features for classifiers trained on data from high-end eye-trackers. The fact that our Webcam-GRU classifier works better on the reading task than on the picture description task suggests that our approach might be more suitable for classification with tasks that involve similar gaze patterns across

participants. Thus, an exciting area of future work will be to investigate if our results generalize to different types of classification tasks and data modalities. With respect to our AD classification task, we plan to investigate different deep learning architectures and different ways to address the issues related to sequence length. In particular, we want to ascertain whether we can use a parameter-heavy transformer architecture by pretraining it on large existing datasets of web-based gaze data collected for mobile devices (e.g., GazeCapture (Krafka et al. 2016)) and then fine-tune it on our data for predicting AD.

Bibliography

Baltrusaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. "OpenFace 2.0: Facial Behavior Analysis Toolkit." Pp. 59–66 in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*.

Barral, Oswald, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. 2020. "Non-Invasive Classification of Alzheimer's Disease Using Eye Tracking and Language." Pp. 813–41 in *Proceedings of the 5th Machine Learning for Healthcare Conference*. PMLR.

Biondi, Juan, Gerardo Fernandez, Silvia Castro, and Osvaldo Agamennoni. 2018. "Eye-Movement Behavior Identification for AD Diagnosis."

Bixler, Robert, and Sidney D'Mello. 2016. "Automatic Gaze-Based User-Independent Detection of Mind Wandering during Computerized Reading." *User Modeling and User-Adapted Interaction* 26(1):33–68. doi: 10.1007/s11257-015-9167-1.

Boote, Bikram, Mansi Agarwal, and Jack Mostow. 2021. "Early Prediction of Children's Disengagement in a Tablet Tutor Using Visual Features." Pp. 98–103 in *Artificial Intelligence in Education, Lecture Notes in Computer Science,* edited by I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova. Cham: Springer International Publishing.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches."

Cordell, Cyndy B., Soo Borson, Malaz Boustani, Joshua Chodosh, David Reuben, Joe Verghese, William Thies, and Leslie B. Fried. 2013. "Alzheimer's Association Recommendations for Operationalizing the Detection of Cognitive Impairment during the Medicare Annual Wellness Visit in a Primary Care Setting." *Alzheimer's & Dementia* 9(2):141–50. doi: 10.1016/j.jalz.2012.09.011.

Cummings, Louise. 2019. "Describing the Cookie Theft Picture: Sources of Breakdown in Alzheimer's Dementia." *Pragmatics and Society* 10(2):153–76. doi: 10.1075/ps.17011.cum.

Fraser, Kathleen C., Jed A. Meltzer, and Frank Rudzicz. 2016. "Linguistic Features Identify Alzheimer's Disease in Narrative Speech." *Journal of Alzheimer's Disease* 49(2):407–22. doi: 10.3233/JAD-150520.

Futoma, Joseph, Sanjay Hariharan, Katherine Heller, Mark Sendak, Nathan Brajer, Meredith Clement, Armando Bedoya, and Cara O'Brien. 2017. "An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection." Pp. 243–54 in *Proceedings of the 2nd Machine Learning for Healthcare Conference*. PMLR.

Garbutt, Siobhan, Alisa Matlin, Joanna Hellmuth, Ana K. Schenk, Julene K. Johnson, Howard Rosen, David Dean, Joel Kramer, John Neuhaus, Bruce L. Miller, Stephen G. Lisberger, and Adam L. Boxer. 2008. "Oculomotor Function in Frontotemporal Lobar Degeneration, Related Disorders and Alzheimer's Disease." *Brain* 131(5):1268–81. doi: 10.1093/brain/awn047.

Goodglass, Harold, and Edith Kaplan. 1972. *The Assessment of Aphasia and Related Disorders*. Lea & Febiger.

Grnarova, Paulina, Florian Schmidt, Stephanie L. Hyland, and Carsten Eickhoff. 2016. "Neural Document Embeddings for Intensive Care Patient Mortality Prediction." doi: 10.48550/arXiv.1612.00467.

Hutt, Stephen, and Sidney K. D'Mello. 2022. "Evaluating Calibration-Free Webcam-Based Eye Tracking for Gaze-Based User Modeling." Pp. 224–35 in *Proceedings of the 2022 International Conference on Multimodal Interaction, ICMI '22*. New York, NY, USA: Association for Computing Machinery.

Jang, Hyeju, Thomas Soroski, Matteo Rizzo, Oswald Barral, Anuj Harisinghani, Sally Newton-Mason, Saffrin Granby, Thiago Monnerat Stutz da Cunha Vasco, Caitlin Lewis, and Pavan Tutt. 2021. "Classification of Alzheimer's Disease Leveraging Multi-Task Machine Learning Analysis of Speech and Eye-Movement Data." *Frontiers in Human Neuroscience* 512.

Kar, Anuradha, and Peter Corcoran. 2017. "A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms." *IEEE Access* 5:16495–519. doi: 10.1109/ACCESS.2017.2735633.

Karlekar, Sweta, Tong Niu, and Mohit Bansal. 2018. "Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models."

Kong, Weirui, Hyeju Jang, Giuseppe Carenini, and Thalia Field. 2019. "A Neural Model for Predicting Dementia from Language." Pp. 270–86 in *Proceedings of the 4th Machine Learning for Healthcare Conference*. PMLR.

Krafka, Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. "Eye Tracking for Everyone." Pp. 2176–84 in.

Lallé, Sébastien, Cristina Conati, and Giuseppe Carenini. 2016. "Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data." Pp. 2529–35 in *IJCAI*.

Li, Shuai, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. "Independently Recurrent Neural Network (Indrnn): Building a Longer and Deeper Rnn." Pp. 5457–66 in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lipton, Zachary C., David C. Kale, Charles Elkan, and Randall Wetzel. 2015. “Learning to Diagnose with LSTM Recurrent Neural Networks.” doi: 10.48550/arXiv.1511.03677.

MacAskill, Michael R., and Tim J. Anderson. 2016. “Eye Movements in Neurodegenerative Diseases.” *Current Opinion in Neurology* 29(1):61–68. doi: 10.1097/WCO.0000000000000274.

Molitor, Robert J., Philip C. Ko, and Brandon A. Ally. 2015. “Eye Movements in Alzheimer’s Disease.” *Journal of Alzheimer’s Disease : JAD* 44(1):1–12. doi: 10.3233/JAD-141173.

Müller, Philipp, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. “Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour.” Pp. 1–10 in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA ’18*. New York, NY, USA: Association for Computing Machinery.

Pavisic, Ivanna M., Nicholas C. Firth, Samuel Parsons, David Martinez Rego, Timothy J. Shakespeare, Keir X. X. Yong, Catherine F. Slattery, Ross W. Paterson, Alexander J. M. Foulkes, Kirsty Macpherson, Amelia M. Carton, Daniel C. Alexander, John Shawe-Taylor, Nick C. Fox, Jonathan M. Schott, Sebastian J. Crutch, and Silvia Primativo. 2017. “Eyetracking Metrics in Young Onset Alzheimer’s Disease: A Window into Cognitive Visual Functions.” *Frontiers in Neurology* 8.

Pusiol, Guido, Andre Esteva, Scott S. Hall, Michael Frank, Arnold Milstein, and Li Fei-Fei. 2016. “Vision-Based Classification of Developmental Disorders Using Eye-Movements.” Pp. 317–25 in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Lecture Notes in Computer Science*, edited by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells. Cham: Springer International Publishing.

Sims, Shane D., and Cristina Conati. 2020. “A Neural Architecture for Detecting User Confusion in Eye-Tracking Data.” Pp. 15–23 in *Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI ’20*. New York, NY, USA: Association for Computing Machinery.

Tran, Minh, Taylan Sen, Kurtis Haut, Mohammad Rafayet Ali, and Ehsan Hoque. 2022. “Are You Really Looking at Me? A Feature-Extraction Framework for Estimating Interpersonal Eye Gaze From Conventional Video.” *IEEE Transactions on Affective Computing* 13(2):912–25. doi: 10.1109/TAFFC.2020.2979440.

Trauzettel-Klosinski, Susanne, Klaus Dietz, and the IReST Study Group. 2012. “Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST.” *Investigative Ophthalmology & Visual Science* 53(9):5452–61. doi: 10.1167/iovs.11-8284.

Tripathi, Subarna, Zachary C. Lipton, Serge Belongie, and Truong Nguyen. 2016. “Context Matters: Refining Object Detection in Video with Recurrent Neural Networks.” doi: 10.48550/arXiv.1607.04648.

Valliappan, Nachiappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, and Vidhya Navalpakkam. 2020. “Accelerating Eye Movement Research via Accurate and Affordable Smartphone Eye Tracking.” *Nature Communications* 11(1):4553. doi: 10.1038/s41467-020-18360-5.

Yue-Hei Ng, Joe, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. “Beyond Short Snippets: Deep Networks for Video Classification.” Pp. 4694–4702 in.

Zhang, Xucong, Yusuke Sugano, and Andreas Bulling. 2017. “Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery.” Pp. 193–203 in *Proceedings of the 30th annual ACM symposium on user interface software and technology*.

Appendix - Detailed information regarding implementations of the GRU networks

Our implementation of the intermediate losses for the GRU network (as shown in Chapter 5.2) was done independent of the fact that Lipton et al 2015 had already introduced Target Replication, hence the implementation was naïve and included manual backpropagation steps after every *k* timesteps in the sequence.

On the other hand, Target Replication as presented in (Lipton et al. 2015) implements a custom loss function, as shown below:

$$\alpha \cdot \frac{1}{T}\sum_{t=1}^{T} loss(\hat{y}^{(t)}, y^{(t)}) \; + \; (1 - \alpha) \cdot \; loss(\hat{y}^{(T)}, y^{(T)})$$

where *T* is the total number of sequence steps, *loss* is a cross-entropy loss function and $\alpha \in [0,1]$ is a hyperparameter which determines the relative importance of intermediate losses.

We completed an experiment using the above loss function for Target Replication at every timestep, however, the accuracy achieved was worse than the Webcam-GRU reported in this work. Moreover, this implementation did not perform better than the majority baseline. Intuitively, we expected Target Replication at every timestep to perform better than Webcam-GRU, which backpropagates errors at every *k*=100 timesteps since the former would perform more updates to the network and hopefully improve performance, however, we see the opposite. We believe that due to lack of enough information in each individual timestep of the sequence with respect to the overall target class of the sequence, i.e., patient or healthy control, the network does not learn to distinguish between the classes. Rather, in the case of backpropagating

after every k=100 timesteps, the network receives information from and backpropagates through sub-sequences of size k=100 and learns to distinguish between classes more effectively.